18 How Genomes Evolve

18.1 Genomes: the First Ten Billion Years

18.2 Acquisition of New Genes

18.3 Noncoding DNA and Genome Evolution

18.4 The Human Genome: the Last Five Million Years

When you have read Chapter 18, you should be able to:

- Explain why biologists believe that the first genomes were made of RNA.
- Describe the events thought to have led to adoption of protein catalysts and DNA genomes by the earliest cells.
- Distinguish between the various ways in which genomes can acquire new genes.
- Give examples to illustrate the role of gene duplication in the evolution of multigene families.
- Discuss the evidence for whole genome duplications in the evolutionary histories of *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and humans.
- Describe the impact that segment duplications have had on the recent evolution of the human genome.
- Explain how new genes can arise by domain duplication and domain shuffling.
- Assess the likely impact of lateral gene transfer on genome evolution in bacteria and in eukaryotes.
- Outline how transposable elements may have influenced genome evolution.
- Define and evaluate the "introns early" and "introns late" hypotheses.
- List the differences between the human and chimpanzee genomes and discuss how such similar genomes can give rise to such different biological attributes.

Mutation and recombination provide the genome with the means to evolve, but we learn very little about the evolutionary histories of genomes simply by studying these events in living cells. Instead we must combine our understanding of mutation and recombination with comparisons between the genomes of different organisms in order to infer the patterns of genome evolution that have occurred. Clearly, this approach is imprecise and uncertain but, as we will see, it is based on a surprisingly large amount of hard data and we can be reasonably confident that, at least in outline, the picture that emerges is not too far from the truth.

In this chapter we will explore the evolution of genomes from the very origins of biochemical systems through to the present day. We will look at ideas regarding the **RNA world**, prior to the appearance of the first DNA molecules, and then examine how DNA genomes have gradually become more complex. Finally, in Section 18.4 we will compare the human genome with the genome of the chimpanzee in order to identify the evolutionary changes that have occurred during the last five million years and which must, somehow, make us what we are.

18.1 Genomes: the First Ten Billion Years

Cosmologists believe that the universe began some 14 billion years ago with the gigantic "primordial fireball" called the Big Bang. Mathematical models suggest that after about 4 billion years galaxies began to fragment from the clouds of gas emitted by the Big Bang, and that within our own galaxy the solar nebula condensed to form the Sun and its planets about 4.6 billion years ago (Figure 18.1). The early Earth was covered with water and it was in this vast planetary ocean that the first biochemical systems appeared, cellular life being well established by the time land masses began to appear, some 3.5 billion years ago. But cellular life was a relatively late stage in biochemical evolution, being preceded by self-replicating polynucleotides that were the progenitors of the first genomes. We must begin our study of genome evolution with these precellular systems.

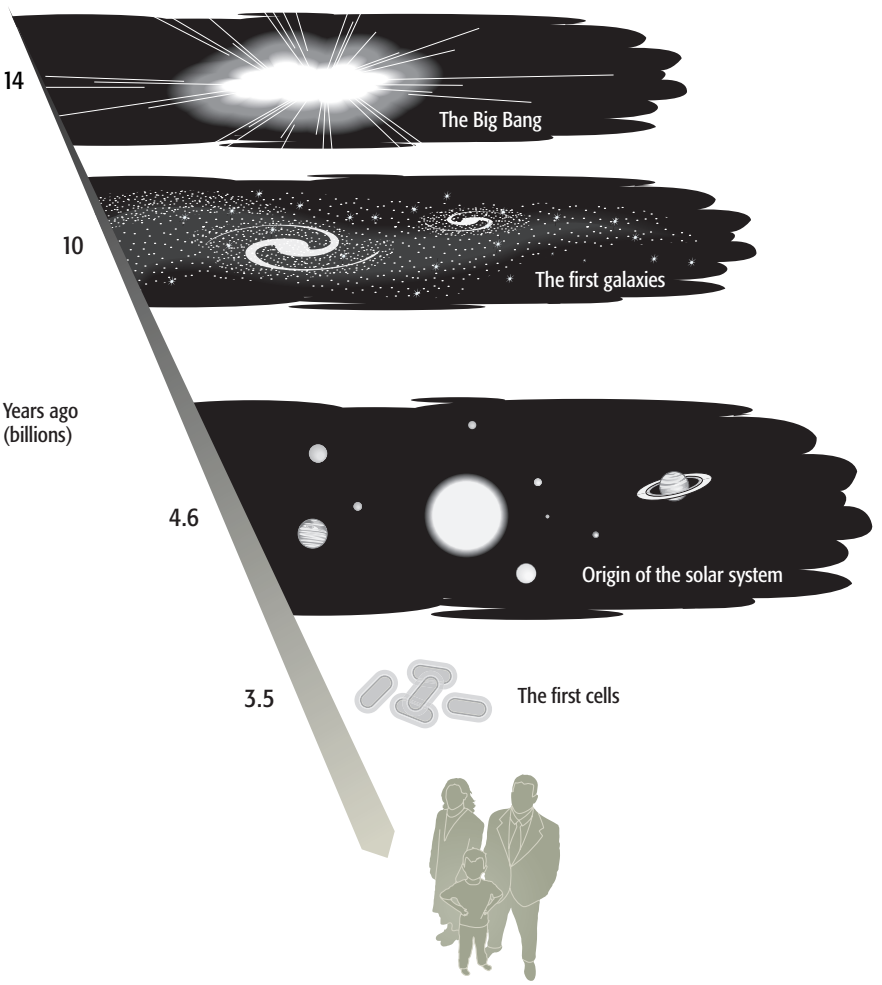


Figure 18.1 The origins of the universe, galaxies, solar system, and cellular life.

18.1.1 The origins of genomes

The first oceans are thought to have had a similar salt composition to those of today but the Earth's atmosphere, and hence the dissolved gases in the oceans, was very different. The oxygen content of the atmosphere remained very low until photosynthesis evolved, and to begin with the most abundant gases were probably methane and ammonia. Experiments attempting to recreate the conditions in the ancient atmosphere have shown that electrical discharges in a methane–ammonia mixture result in chemical synthesis of a range of amino acids, including alanine, glycine, valine, and several of the others found in proteins. Hydrogen cyanide and formaldehyde are also formed, these participating in additional reactions to give other amino acids, as well as purines, pyrimidines, and, in less abundance, sugars. At least some of the building blocks of biomolecules could therefore have accumulated in the ancient chemosphere.

The first biochemical systems were centered on RNA

Polymerization of the building blocks into biomolecules might have occurred in the oceans or could have been promoted by the repeated condensation and drying of droplets of water in clouds. Alternatively, polymerization might have taken place on solid surfaces, perhaps making use of monomers immobilized on clay particles, or in hydrothermal vents. The precise mechanism need not concern us: what is important is that it is possible to envisage purely geochemical processes that could lead to synthesis of polymeric biomolecules similar to the ones found in living systems. It is the next steps that we must worry about. We have to go from a random collection of biomolecules to an ordered assemblage that displays at least some of the biochemical properties that we associate with life. These steps have never been reproduced experimentally and our ideas are therefore based mainly on speculation tempered by a certain amount of computer simulation. One problem is that the speculations are unconstrained because the global ocean could have contained as many as 10^{10} biomolecules per liter and we can allow a billion years for the necessary events to take place. This means that even the most improbable scenarios cannot be dismissed out of hand.

Progress in understanding the origins of life was initially stalled by the apparent requirement that polynucleotides and polypeptides must work in harness in order to produce a self-reproducing biochemical system. This is because proteins are required to catalyze biochemical reactions but cannot carry out their own self-replication. Polynucleotides can specify the synthesis of proteins and self-replicate, but it was thought that they could do neither without the aid of proteins. It appeared that the biochemical system would have to spring fully formed from the random collection of biomolecules because any intermediate stage could not be perpetuated. The major breakthrough came in the mid-1980s when it was discovered that RNA can have catalytic activity.

Those ribozymes that are known today carry out three types of biochemical reaction:

- Self-cleavage, as displayed by the self-splicing Group I, II, and III introns and by some virus genomes (Table 12.4 and Section 12.2.4).
- Cleavage of other RNAs, as carried out by, for example, RNase P (Table 12.4 and Section 12.1.3).
- Synthesis of peptide bonds, carried out by the rRNA component of the ribosome (Section 13.2.3).

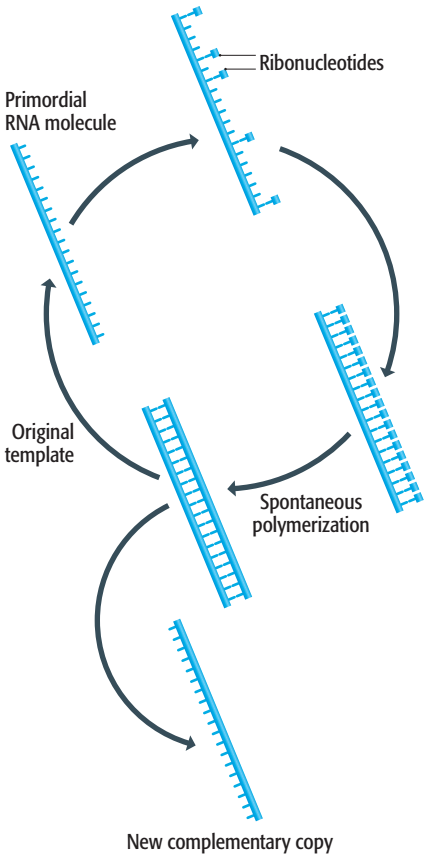


Figure 18.2 Copying of RNA molecules in the early RNA world. Before the evolution of RNA polymerases, ribonucleotides that became associated with an RNA template would have had to polymerize spontaneously. This process would have been inaccurate and many RNA sequences would have been generated

In the test tube, synthetic RNA molecules have been shown to carry out other biologically relevant reactions such as synthesis of ribonucleotides, synthesis and copying of RNA molecules, and transfer of an RNA-bound amino acid to a second amino acid forming a dipeptide, in a manner analogous to the role of tRNA in protein synthesis. The discovery of these catalytic properties solved the polynucleotide–polypeptide dilemma by showing that the first biochemical systems could have been centered entirely on RNA.

Ideas about the RNA world have taken shape in recent years. We now envisage that RNA molecules initially replicated in a slow and haphazard fashion simply by acting as templates for binding of complementary nucleotides which polymerized spontaneously (Figure 18.2). This replication process would have been very inaccurate so a variety of RNA sequences would have been generated, eventually leading to one or more with nascent ribozyme properties that were able to direct their own, more accurate self-replication. It is possible that a form of natural selection operated so that the most efficient replicating systems began to predominate, as has been shown to occur in experimental systems. A greater accuracy in replication would have enabled RNAs to increase in length without losing their sequence specificity, providing the potential for more sophisticated catalytic properties, possibly culminating in structures as complex as present-day Group I introns (see Figure 12.39) and ribosomal RNAs (see Figure 13.11).

To call the early RNAs "genomes" is a little fanciful, but the term **protogenome** has attractions as a descriptor for molecules that were self-replicating and able to direct simple biochemical reactions. These reactions might have included energy metabolism based, as today, on the release of free energy by hydrolysis of the phosphate–phosphate bonds in the ribonucleotides ATP and GTP, and the reactions might have become compartmentalized within lipid membranes, forming the first cell-like structures. There are difficulties in envisaging how long-chain unbranched lipids could form by chemical- or ribozyme-catalyzed reactions, but once present in sufficient quantities they would have assembled spontaneously into membranes, possibly encapsulating one or more protogenomes and providing the RNAs with an enclosed environment in which more controlled biochemical reactions could be carried out.

The first DNA genomes

How did the RNA world develop into the DNA world? The first major change was probably the development of protein enzymes, which supplemented, and eventually replaced, most of the catalytic activities of ribozymes. There are several unanswered questions relating to this stage of biochemical evolution, including the reason why the transition from RNA to protein occurred in the first place. Originally, it was assumed that the 20 amino acids in polypeptides provided proteins with greater chemical variability than the four ribonucleotides in RNA, enabling protein enzymes to catalyze a broader range of biochemical reactions, but this explanation has become less attractive as more and more ribozyme-catalyzed reactions have been demonstrated in the test tube. A more recent suggestion is that protein-mediated catalysis is more efficient because of the inherent flexibility of folded polypeptides compared with the greater rigidity of base-paired RNAs. Alternatively, enclosure of RNA protogenomes within membrane vesicles

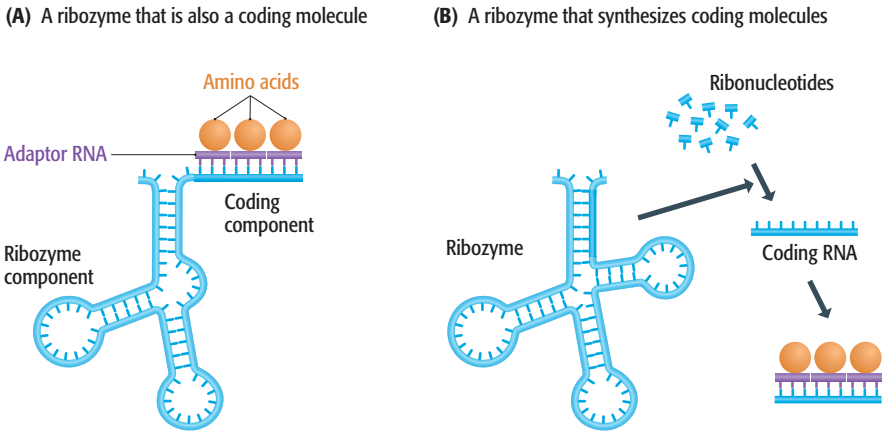


Figure 18.3 **Two scenarios for the evolution of the first coding RNA.** A ribozyme could have evolved to have a dual catalytic and coding function (A), or a ribozyme could have synthesized a coding molecule (B). In both examples, the amino acids are shown attaching to the coding molecule via small adaptor RNAs, the presumed progenitors of today's tRNAs.

could have prompted the evolution of the first proteins, because RNA molecules are hydrophilic and must be given a hydrophobic coat, for instance by attachment to peptide molecules, before being able to pass through or become integrated into a membrane.

The transition to protein-mediated catalysis demanded a radical shift in the function of the RNA protogenomes. Rather than being directly responsible for the biochemical reactions occurring in the early cell-like structures, the protogenomes became coding molecules whose main function was to specify the construction of the catalytic proteins. Whether the ribozymes themselves became coding molecules, or coding molecules were synthesized by the ribozymes is not known, although the most persuasive theories about the origins of protein synthesis and the genetic code suggest that the latter alternative is more likely to be correct (Figure 18.3). Whatever the mechanism, the result was the paradoxical situation whereby the RNA protogenomes had abandoned their roles as enzymes, which they were good at, and taken on a coding function for which they were less well suited because of the relative instability of the RNA phosphodiester bond (Section 1.2.1). A transfer of the coding function to the more stable DNA seems almost inevitable and would not have been difficult to achieve, reduction of ribonucleotides giving deoxyribonucleotides which could then be polymerized into copies of the RNA protogenomes by a reverse-transcriptase-catalyzed reaction (Figure 18.4). The replacement of uracil with its methylated derivative thymine probably conferred even more stability on the DNA polynucleotide, and the adoption of double-stranded DNA as the coding molecule was almost certainly prompted by the possibility of repairing DNA damage by copying the partner strand (Sections 16.2.2 and 16.2.3).

According to this scenario, the first DNA genomes comprised many separate molecules, each specifying a single protein and each therefore equivalent to a single gene. The linking together of these genes into the first chromosomes, which could have begun before the transition to DNA, would have improved the efficiency of gene distribution during cell division, as it is easier to organize the equal distribution of a few large chromosomes than many separate genes. As with most stages in early genome evolution, several different mechanisms by which genes might have become linked have been proposed.

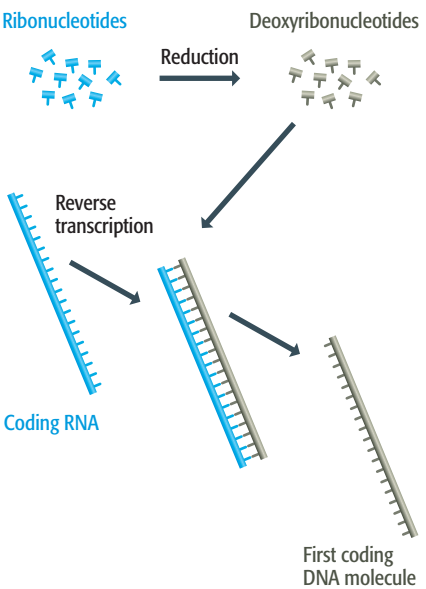


Figure 18.4 **Conversion of a coding RNA molecule into the progenitor of the first DNA genome.**

How unique is life?

If the experimental simulations and computer models are correct then it is likely that the initial stages in biochemical evolution occurred many times in parallel in the oceans or atmosphere of the early Earth. It is therefore quite possible that "life" arose on more than one occasion, even though all present-day organisms appear to derive from a single origin. This single origin is indicated by the remarkable similarity between the basic molecular biological and biochemical mechanisms in bacterial, archaeal, and eukaryotic cells. To take just one example, there appears to be no obvious biological or chemical reason why any particular triplet of nucleotides should code for any particular amino acid, but the genetic code, although not universal, is virtually the same in all organisms that have been studied. If these organisms derived from more than one origin then we would anticipate two or more very different codes.

If multiple origins are possible, but modern life is derived from just one origin, then at what stage did a single biochemical system begin to predominate? The question cannot be answered precisely, but the most likely scenario is that the predominant system was the first to develop the means to synthesize protein enzymes and therefore probably also the first to adopt a DNA genome. The greater catalytic potential and more accurate replication conferred by protein enzymes and DNA genomes would have given these cells a significant advantage compared with those still containing RNA protogenomes. The DNA–RNA–protein cells would have multiplied more rapidly, enabling them to out-compete the RNA cells for nutrients which, before long, would have included the RNA cells themselves.

Are life forms based on informational molecules other than DNA and RNA possible? It is not impossible that RNA was preceded by some other informational molecule at the very earliest period of biochemical evolution. In particular, a pyranosyl version of RNA, in which the sugar takes on a slightly different structure, might be a better choice than normal RNA for an early protogenome because the base-paired molecules that it forms are more stable. The same is true of **peptide nucleic acid** (**PNA**), a polynucleotide analog in which the sugar–phosphate backbone is replaced by amide bonds (Figure 18.5). PNAs have been synthesized in the test tube and have been shown to form base pairs with normal polynucleotides. However, there are no indications that either pyranosyl RNA or PNA were more likely than RNA to have evolved in the prebiotic soup.

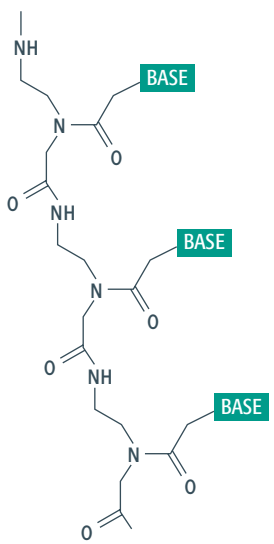


Figure 18.5 A short stretch of peptide nucleic acid. A peptide nucleic acid has an amide backbone instead of the sugar–phosphate structure found in a standard nucleic acid.

18.2 Acquisition of New Genes

Although the very old fossil record is difficult to interpret, there is reasonably convincing evidence that by 3.5 billion years ago biochemical systems had evolved into cells similar in appearance to modern bacteria. We cannot tell from the fossils what kinds of genomes these first real cells had, but from the discussion in the preceding section we can infer that they were made of double-stranded DNA and consisted of a small number of chromosomes, possibly just one, each containing many linked genes.

If we follow the fossil record forward in time we see the first evidence for eukaryotic cells—structures resembling single-celled algae—about 1.4 billion years ago (Figure 18.6) and for multicellular algae about 0.9 billion years ago. Multicellular animals appeared around 640 million years ago, although

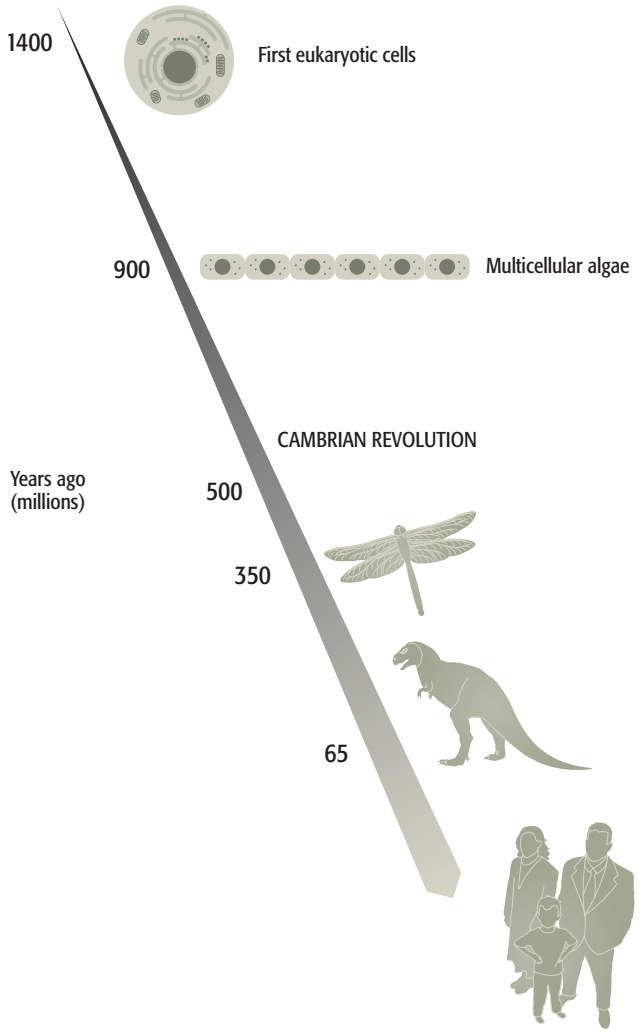


Figure 18.6 The evolution of life.

there are enigmatic burrows suggesting that animals lived earlier than this. The Cambrian Revolution, when invertebrate life proliferated into many novel forms, occurred 530 million years ago and ended with the disappearance of many of the novel forms in a mass extinction 500 million years ago. Since then, evolution has continued apace and with increasing diversification: the first terrestrial insects, animals, and plants were established by 350 million years ago, the dinosaurs had been and gone by the end of the Cretaceous, 65 million years ago, and the first hominoids appeared a mere 4.5 million years ago.

Morphological evolution was accompanied by genome evolution. It is dangerous to equate evolution with "progress" but it is undeniable that as we move up the evolutionary tree we see increasingly complex genomes. One indication of this complexity is gene number, which varies from less than 1000 in some bacteria to 30,000–40,000 in vertebrates such as humans. Within individual lineages—within the bacteria, for example—changes in gene number have probably been gradual, with the acquisition of new genes balanced at least in part by the loss of existing ones. We should also remember that the evolutionary pathways of some organisms have involved a *decrease* rather than increase in gene number, leading to the minimal genomes of

Mycoplasma and other parasitic species (Section 8.2.2). The gradual changes within lineages were, however, punctuated by at least two transition periods when new organisms with greatly increased gene numbers appeared. One of these transitions accompanied the arrival of the first eukaryotes about 1.4 billion years ago, these cells probably containing at least 10,000 genes (the minimum in modern eukaryotes) compared with the 5000 or fewer typical of prokaryotes. The second transition was associated with the arrival of the first vertebrates soon after the end of the Cambrian, these, like modern vertebrates, probably having at least 30,000 genes.

There are two fundamentally different ways in which new genes could be acquired by a genome:

- By duplicating some or all of the existing genes in the genome.
- By acquiring genes from other species.

Both events have been important in genome evolution, as we will see in the next two sections.

18.2.1 Acquisition of new genes by duplication events

A central role for gene duplication in genome evolution was first proposed in 1970. The initial result of gene duplication will be two identical genes. Selective constraints will ensure that one of these genes retains its original nucleotide sequence, or something very similar to it, so that it can continue to provide the protein function that was originally supplied by the single gene copy before the duplication took place. It is possible that the same selective constraints will apply to the second gene, especially if the increase in the rate of synthesis of the gene product, made possible by the duplication, confers a benefit on the organism (Figure 18.7). More frequently, however, the second copy will confer no benefit and hence will not be subject to the same selective pressures and so will accumulate mutations at random. Evidence shows that the majority of new genes that arise by duplication acquire deleterious mutations that inactivate them so that they become pseudogenes, examination of existing pseudogenes suggesting that the commonest inactivating mutations are frameshifts and nonsense mutations that occur within the coding region of the gene. Occasionally, however, mutations might not lead to inactivation of the gene but instead result in a new gene function that is useful to the organism (see Figure 18.7).

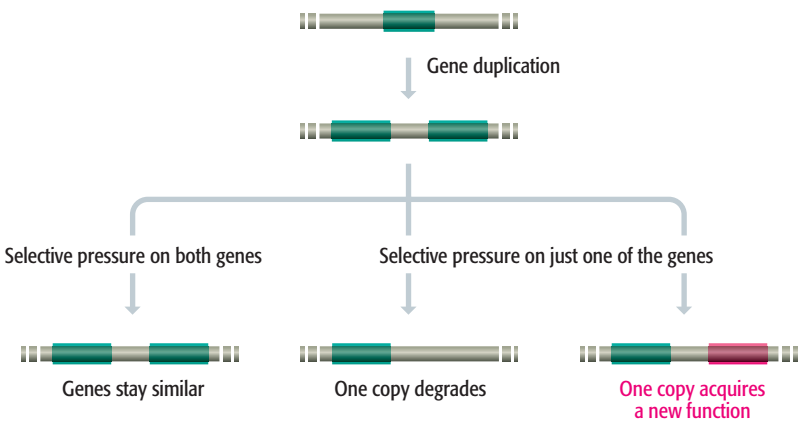


Figure 18.7 Three scenarios for the outcome of a gene duplication.

We will first consider the evidence for past gene duplications that is contained in present-day genome sequences, and then look at the mechanisms by which gene duplication can occur.

Genome sequences provide extensive evidence of past gene duplications

The most cursory examination of a genome sequence provides ample evidence that many genes have arisen by duplication events. The importance of the first scenario illustrated in Figure 18.7, where the increased amount of gene product resulting from a gene duplication is beneficial and stabilizes the duplication, is supported by the many examples of multigene families made up of genes with identical or near-identical sequences. The prime examples are the rRNA genes, whose copy numbers range from two in *Mycoplasma genitalium* to 500 or more in *Xenopus laevis*, with all of the copies having virtually the same sequence. These multiple copies of identical genes presumably reflect the need for rapid synthesis of rRNAs at certain stages of the cell cycle. Note that the existence of these multigene families indicates not only that gene duplications have occurred in the past but also that there must be a molecular mechanism that ensures that the family members retain their identity over evolutionary time. This is called **concerted evolution**. If one copy of the family acquires an advantageous mutation then it is possible for that mutation to spread throughout the family until all members possess it. The most likely way in which this can be achieved is by gene conversion which, as described in Section 17.1.1, can result in the sequence of one copy of a gene being replaced with all or part of the sequence of a second copy. Multiple gene conversion events could therefore maintain identity among the sequences of the individual members of a multigene family, especially if those members are arranged in tandem arrays.

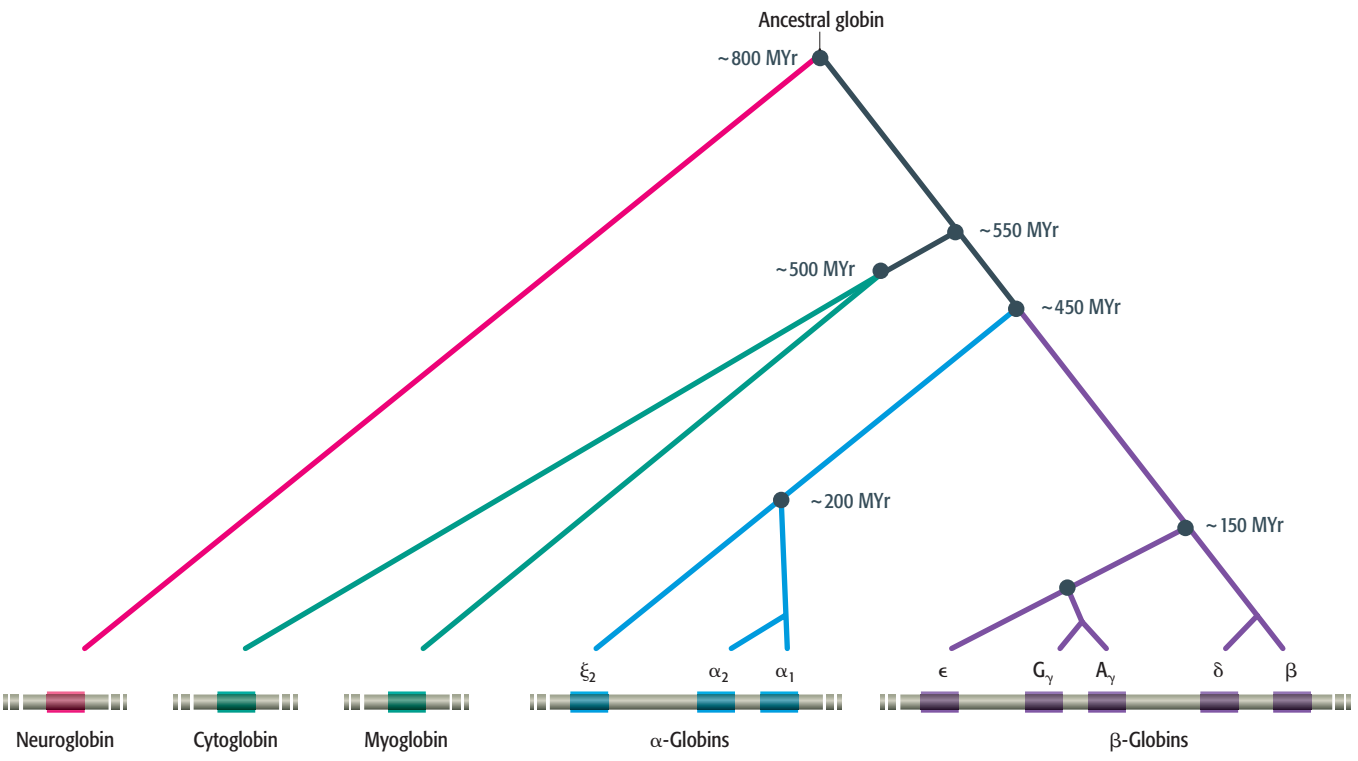


Figure 18.8 Evolution of the globin gene superfamily of humans. The members of the superfamily are now on different chromosomes. The neuroglobin gene is on chromosome 14, the cytoglobin gene is on chromosome 17, and the myoglobin gene is on chromosome 22. The α-globin cluster is on chromosome 16 and the β-globin cluster is on chromosome 11. Abbreviation: MYr, millions of years ago.

The third scenario in Figure 18.7 results in the duplicated gene accumulating mutations that give it a new, useful function. Again, multigene families provide many indications that such events have occurred frequently in the past. We have already seen that gene duplications in the globin gene families led to the evolution of new globin proteins that are used by the organism at different stages in its development (see Figure 7.19). We noted that all the globin genes, both the α- and β-types, have related nucleotide sequences and hence form part of a superfamily that includes genes specifying various other proteins which, like the blood globins, have the capacity to bind oxygen molecules. From the degrees of similarity displayed by pairs of genes in the superfamily, it is possible to deduce the pattern of gene duplications that gave rise to the genes we see today, and by applying the **molecular clock** (Section 19.2.2) to the data we can estimate how many millions of years ago each duplication took place. These analyses tell us that a duplication some 800 million years ago resulted in a pair of ancestral genes, one of which evolved into the modern gene for the brain protein neuroglobin, and the other of which gave rise to all the other members of the superfamily (Figure 18.8). Some 250 million years later there was a second duplication on the path leading to the blood globins, one of the products of this duplication being a gene that, via another duplication, gave rise to myoglobin, which is active in muscle, and cytoglobin, which is present in many tissues but whose function is not yet understood. The proto-α and proto-β lineages split by a duplication that occurred 450 million years ago and the duplications within the α- and β-globin gene families took

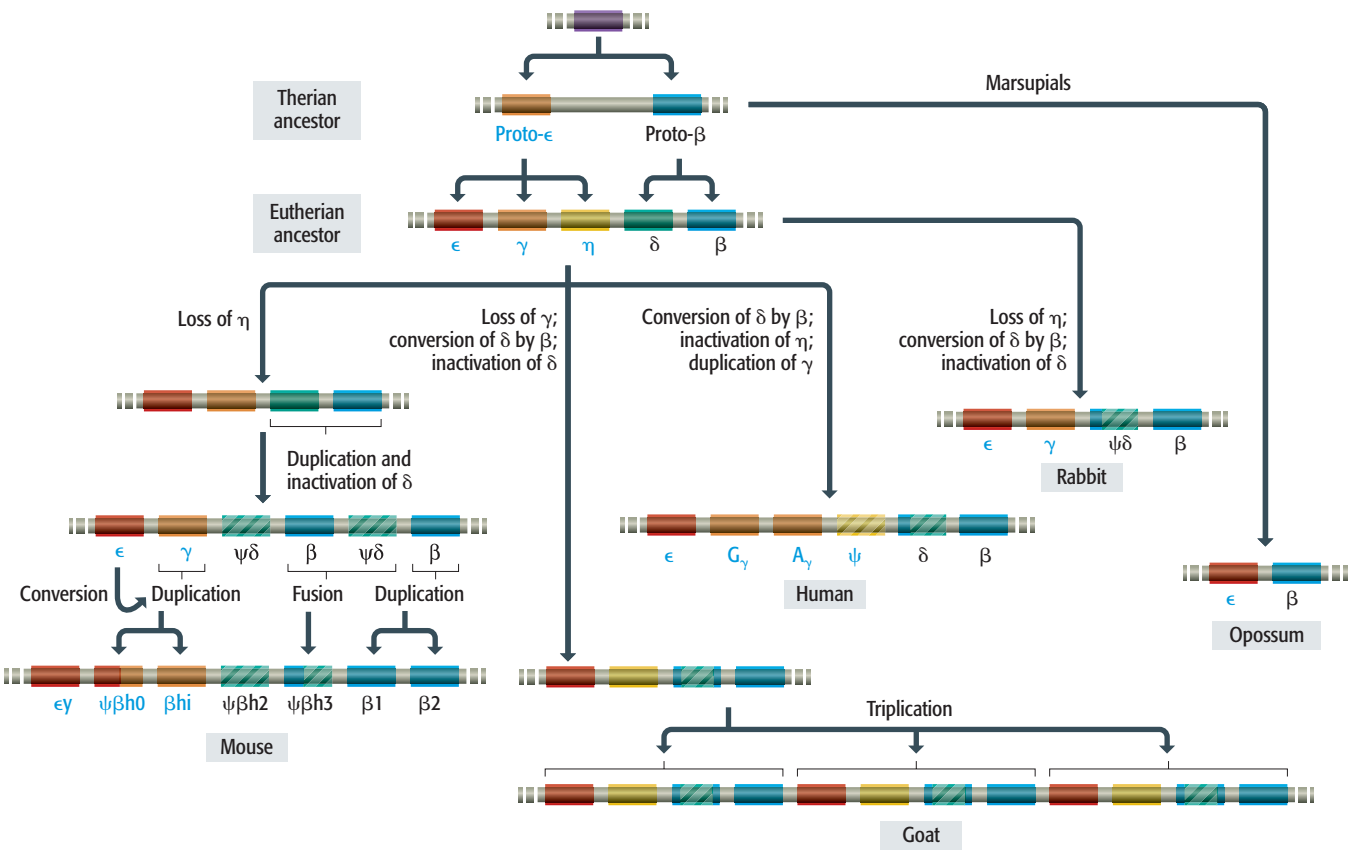


Figure 18.9 The evolution of the mammalian β-globin genes. Reprinted from *Genomics*, Vol. 13, Tagle *et al.*, 'The β-globin gene cluster …,' 741–760, 1992, wither permission from Elsevier.

place during the last 200 million years. Within this more recent time frame it is possible to deduce not only the pattern of gene duplication but also some of the more detailed changes that have occurred within individual genes. Hence the events leading to the various groups of β-globin genes present in different mammals have been inferred (Figure 18.9).

We observe similar patterns of evolution when we compare the sequences of other genes. The trypsin and chymotrypsin genes, for example, are related by a common ancestral gene that duplicated approximately 1500 million years ago. Both now code for proteases involved in protein breakdown in the vertebrate digestive tract, trypsin cutting other proteins at arginine and lysine amino acids, and chymotrypsin cutting at phenylalanines, tryptophans, and tyrosines. Genome evolution has therefore produced two complementary protein functions where originally there was just one.

Another striking example of gene evolution by duplication is provided by the homeotic selector genes, the key developmental genes responsible for specification of the body plans of animals. As described in Section 14.3.4, *Drosophila* has a single cluster of homeotic selector genes, called HOM-C, which consists of eight genes each containing a homeodomain sequence coding for a DNA-binding motif in the protein product (see Figure 14.37). These eight genes, as well as other homeodomain genes in *Drosophila*, are believed to have arisen by a series of gene duplications that began with an ancestral gene that existed about 1000 million years ago. The functions of the modern genes, each specifying the identity of a different segment of the fruit fly, gives us a tantalizing glimpse of how gene duplication and sequence divergence could, in this case, have been the underlying processes responsible for increasing the morphological complexity of the series of organisms in the *Drosophila* evolutionary tree. If we then move further up the evolutionary tree we see that vertebrates have four Hox gene clusters (see Figure 14.37), each a recognizable copy of the *Drosophila* cluster, with sequence similarities between genes in equivalent positions. The implication is that in the vertebrate lineage there were two duplications, not of individual Hox genes but of the entire cluster (Figure 18.10). Not all of the vertebrate Hox genes have been ascribed functions, but we believe that the additional versions possessed by vertebrates relate to the added complexity of the vertebrate body plan. Two observations support this conclusion. The amphioxus, an invertebrate that displays some primitive vertebrate features, has two Hox clusters, which is what we might expect for a primitive "protovertebrate." Ray-finned fishes, probably the most diverse group of vertebrates with a vast range of different variations of the basic body plan, have seven Hox clusters.

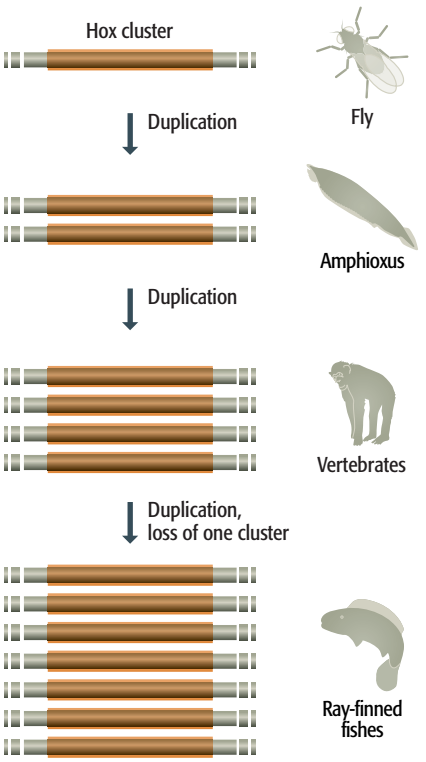


Figure 18.10 Evolution of the Hox gene cluster from flies to ray-finned fishes.

A variety of processes could result in gene duplication

There are several ways in which a short segment of a DNA molecule, possibly containing one or a small group of genes, could be duplicated. These include:

- **Unequal crossing-over**, which is a recombination event that is initiated by similar nucleotide sequences that are not at identical places in a pair of homologous chromosomes. As shown in Figure 18.11A, the result of unequal crossing-over can be duplication of a segment of DNA in one of the recombination products.
- **Unequal sister chromatid exchange**, which occurs by the same mechanism as unequal crossing-over, but involves a pair of chromatids from a single chromosome (Figure 18.11B).

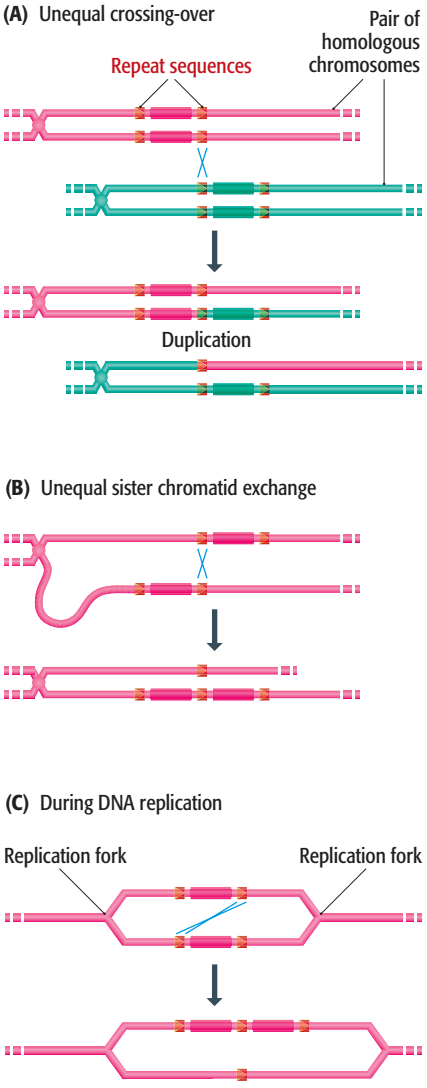


Figure 18.11 Models for gene duplication by (A) unequal crossing-over between homologous chromosomes, (B) unequal sister chromatid exchange, and (C) during replication of a bacterial genome. In each case, recombination occurs between two different copies of a short repeat sequence, leading to duplication of the sequence between the repeats. Unequal crossing-over and unequal sister chromatid exchange are essentially the same except that the first involves chromatids from a pair of homologous chromosomes and the second involves chromatids from a single chromosome. In (C), recombination occurs between two daughter double helices that have just been synthesized by DNA replication.

- DNA amplification, which is sometimes used in this context to describe duplication of segments of DNA in bacteria and other haploid organisms, the duplications arising by unequal recombination between the two daughter DNA molecules in a replication bubble (Figure 18.11C).
- Replication slippage (see Figure 16.5), which can result in duplication of short segments such as microsatellite sequences. This process could conceivably duplicate a region large enough to contain an entire gene, but in practice this is unlikely.

Each of the four processes listed above lead to tandem duplications—ones in which the two duplicated segments lie adjacent to one another in the genome. This is the pattern seen with many multigene families, such as the α-globin gene family on human chromosome 16 and the β-globin family on chromosome 11, but it is not the only possibility. Family members are not always colocated: for example, in the human genome there are three functional genes for the metabolic enzyme aldolase, each on a different chromosome. These copies might have once been present as a tandem array, and become dispersed as part of large-scale genome reorganizations, but it is also possible that the distant locations are a consequence of the duplication process. This would be the case if the duplication occurred by retrotransposition in a manner similar to that thought to lead to formation of processed pseudogenes (see Figure 7.20). A processed pseudogene arises when the mRNA copy of a gene is converted into cDNA and reinserted into the genome. The resulting structure is a pseudogene because it lacks a promoter sequence, this being absent from the mRNA. The pseudogene could conceivably be inserted adjacent to the promoter of an existing gene and hence become active by subverting this promoter for its own use. Gene duplicates that arise in this way are called **retrogenes** and, again, many examples are known in different genomes (in the human genome the testis-specific version of the pyruvate dehydrogenase gene is a retrogene). A distinctive feature of a retrogene is that it lacks any introns present in the parent copy of the gene, as these are not present in the mRNA. Recently it has been realized that complete copies of genes, including not only the introns but also some or all of the promoter sequences, can also be made by reverse transcription. This can occur when the RNA that is reverse-transcribed is not an mRNA, but an antisense copy of the gene made by transcription of the "wrong" polynucleotide (Figure 18.12). We are beginning to realize that antisense RNAs are not uncommon and may indeed have a role in gene regulation, similar perhaps to the mode of action of microRNAs (Section 12.2.6). If converted into cDNA, then an antisense RNA could provide a full-length, functional gene duplicate for insertion into the genome, at a position that could be distant from the original copy.

Whole genome duplication is also possible

The processes described above give rise to relatively short DNA duplications, perhaps a few tens of kilobases in length. Are larger duplications possible? It seems unlikely that duplication of entire chromosomes has played any major role in genome evolution, because we know that duplication of individual human chromosomes, resulting in a cell that contains three copies of one chromosome and two copies of all the others (the condition called **trisomy**), is either lethal or results in a genetic disease, such as Down syndrome, and similar harmful effects have been observed in artificially generated trisomic mutants of *Drosophila*. Probably, the resulting increase in copy numbers for

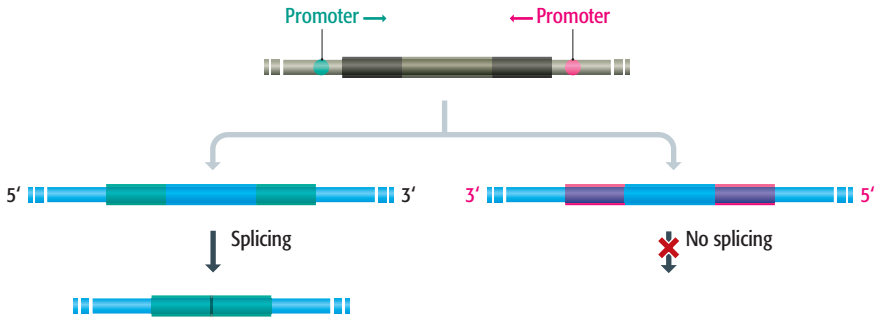


Figure 18.12 An antisense RNA copy of a gene will retain the gene's introns. On the left, the gene is transcribed from its standard promoter to give a pre-mRNA whose intron is removed by splicing. On the right, an antisense RNA is synthesized from a downstream promoter. This RNA is not spliced because the sequences at its exon–intron junctions are not the standard ones recognized by the spliceosome (Section 12.2.2).

some genes but not others leads to an imbalance of the gene products and disruption of the cellular biochemistry.

The harmful effects of trisomy do not mean that duplication of the entire set of chromosomes in a nucleus must be discounted. Genome duplication can occur if an error during meiosis leads to the production of gametes that are diploid rather than haploid (Figure 18.13). If two diploid gametes fuse then the result will be a type of **autopolyploid**, in this case a tetraploid cell whose nucleus contains four copies of each chromosome. Autopolyploidy, as with other types of polyploidy (Section 18.2.2), is not uncommon, especially among plants. Autopolyploids are often viable because each chromosome still has a homologous partner and so can form a bivalent during meiosis. This allows an autopolyploid to reproduce successfully, but generally prevents interbreeding with the original organism from which it was derived. This is because a cross between, for example, a tetraploid and diploid would

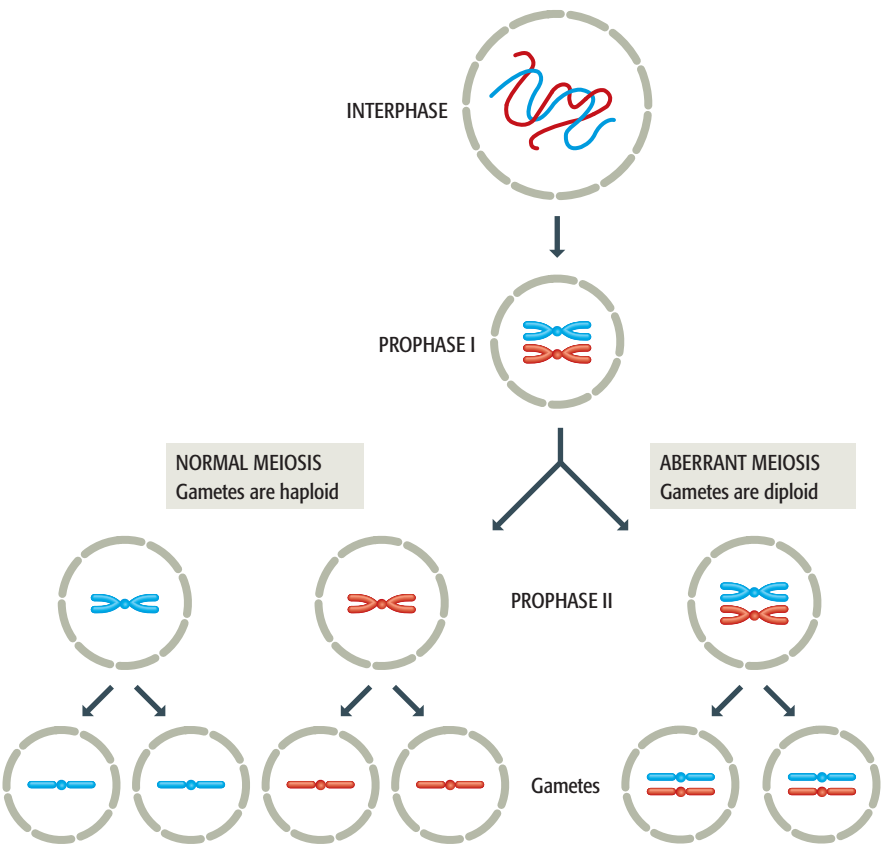


Figure 18.13 The basis of autopolyploidization. The normal events occurring during meiosis are shown, in abbreviated form, on the left (compare with Figure 3.16). On the right, an aberration has occurred between prophase I and prophase II and the pairs of homologous chromosomes have not separated into different nuclei. The resulting gametes will be diploid rather than haploid.

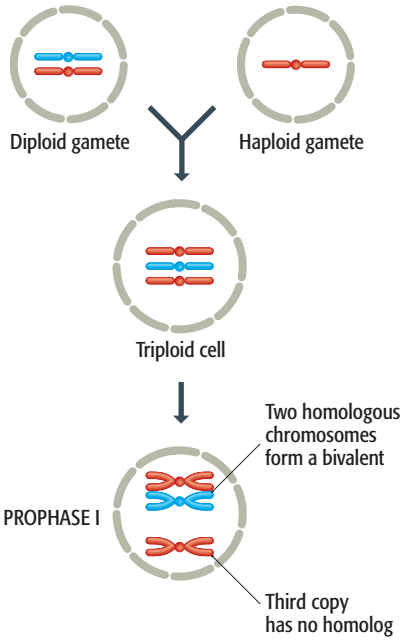


Figure 18.14 Autopolyploids cannot interbreed successfully with their parents. Fusion of a diploid gamete produced by the aberrant meiosis shown in Figure 18.13 with a haploid gamete produced by normal meiosis leads to a triploid nucleus, one that has three copies of each homologous chromosome. During prophase I of the next meiosis, two of these homologous chromosomes will form a bivalent but the third will have no partner. This has a disruptive effect on the segregation of chromosomes during anaphase (see Figure 3.16) and usually prevents meiosis from reaching a successful conclusion. This means that gametes are not produced and the triploid organism is sterile. Note that the bivalent could have formed between any two of the three homologous chromosomes, not just between the pair shown in the diagram.

give a triploid offspring which would not itself be able to reproduce because one full set of its chromosomes would lack homologous partners (Figure 18.14). Autopolyploidy is therefore a mechanism by which speciation can occur, a pair of species usually being defined as two organisms that are unable to interbreed. The generation of new plant species by autopolyploidy has in fact been observed, notably by Hugo de Vries, one of the rediscoverers of Mendel's experiments. During his work with evening primrose, *Oenothera lamarckiana*, de Vries isolated a tetraploid version of this normally diploid plant, which he named *Oenothera gigas*. Autopolyploidy among animals is less common, especially in those with two distinct sexes, possibly because of problems that arise if a nucleus possesses more than one pair of sex chromosomes. It is not impossible though, and at least one mammal, the red viscacha rat of Argentina, has a tetraploid genome.

Analysis of modern genomes provides evidence for past genome duplications

Autopolyploidy does not lead directly to an increase in gene number because the initial product is an organism that simply has extra copies of every gene, rather than any new genes. It does, however, provide the *potential* for an increase because the extra genes are not essential to the functioning of the cell and so can undergo mutational change without harming the viability of the organism. With this in mind, is there any evidence that whole genome duplication has been important in the large-scale acquisition of new genes during the evolutionary histories of present-day genomes?

From what we understand about the way in which genomes change over time, we might anticipate that evidence for whole genome duplication would be quite difficult to obtain. Many of the extra gene copies resulting from genome duplication would decay to the extent that they are no longer visible in the DNA sequence. Those genes that are retained, because their duplicated function is useful to the organism or because they have evolved new functions, should be identifiable, but it would be difficult to distinguish if they have arisen by duplication of the entire genome or by duplication of much smaller segments. For a genome duplication to be signaled it would be necessary to find large, duplicated *sets* of genes, with the same order of genes in both sets. To what extent these duplicated sets are still visible in the genome will depend on how frequently past recombination events have moved genes to new positions.

To search the *Saccharomyces cerevisiae* genome for evidence of past duplications, homology analysis (Section 5.2.1) was carried out with every yeast gene tested against every other yeast gene. To be considered descendants of a duplication event, two genes had to display at least 25% identity when the predicted amino acid sequences of their protein products were compared. About 800 gene pairs were identified in this way, 376 of which could be placed in 55 duplicate sets, each of these sets containing at least three genes in the same order (Figure 18.15), possibly with other genes interspersed between them, the sets altogether covering half the genome. These sets could have arisen by duplication of segments rather than the entire genome, but if this was the case then it might be anticipated that some of the genes would have been duplicated more than once. The fact that there were just two copies of each gene, and never three or four, therefore supported the notion that the copies arose by whole genome duplication. This possibility became more certain when the

complete genome sequences of other yeast species were obtained. Comparisons between the genomes of *S. cerevisiae, Kluyveromyces lactis,* and *Ashbya gossypii* have been particularly informative. These three species shared a common ancestor that lived over 100 million years ago, previous to the time of the genome duplication event inferred from the homology analysis. If that duplication had indeed occurred in the lineage leading to *S. cerevisiae* then it would be anticipated that this species would have duplicated copies of many genes present as singletons in the *K. lactis* and *A. gossypii* genomes. This turns out to be the case, this new analysis suggesting that some 10% of the genes in the modern *S. cerevisiae* genome derive from a whole genome duplication that occurred just under 100 million years ago.

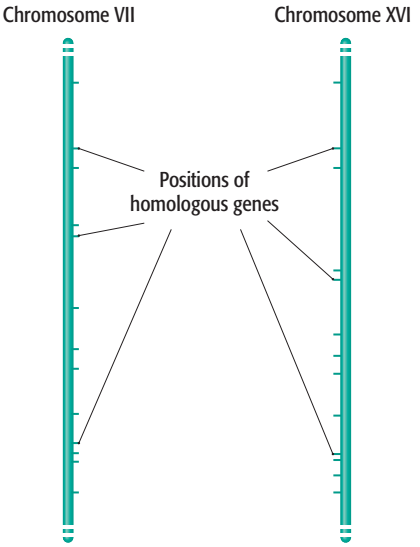


Figure 18.15 An example of a duplicated set of genes in the *S. cerevisiae* genome. Each of the three pairs of genes that are indicated have high sequence homology. This and other duplicated sets provide evidence for a genome duplication just under 100 million years ago in the *S. cerevisiae* lineage.

Equivalent work has been carried out with other genomes and the picture that is emerging is that whole genome duplication has been a relatively frequent event in the evolution of many groups of organisms. For example, comparisons between the *Arabidopsis thaliana* genome sequence and segments of other plant genomes suggest that the ancestor of the *A. thaliana* genome underwent four rounds of genome duplication between 100 million and 200 million years ago. The human genome and other mammalian genomes also contain so many gene duplicates that at least one whole genome duplication event is thought to have occurred in this lineage between 350 million and 600 million years ago.

Smaller duplications can also be identified in the human genome and other genomes

Although the most recent whole genome duplication in the human lineage occurred a considerable time ago, the human genome has not been quiescent in the intervening period. Indeed, the opposite is true, one surprise provided by the human genome sequence being the realization that there has been extensive and frequent duplication of short segments of the genome in the relatively recent past. This is illustrated by Figure 18.16, which depicts the duplication events that are inferred for the last 35 million years of evolution of the long arm of human chromosome 22. As in the yeast study described above, these duplications have been identified by comparing different parts of the genome, in this case to identify regions of more than 1 kb in length that display 90% or higher nucleotide sequence similarity. This analysis ignores interspersed repeat sequences and so locates regions that

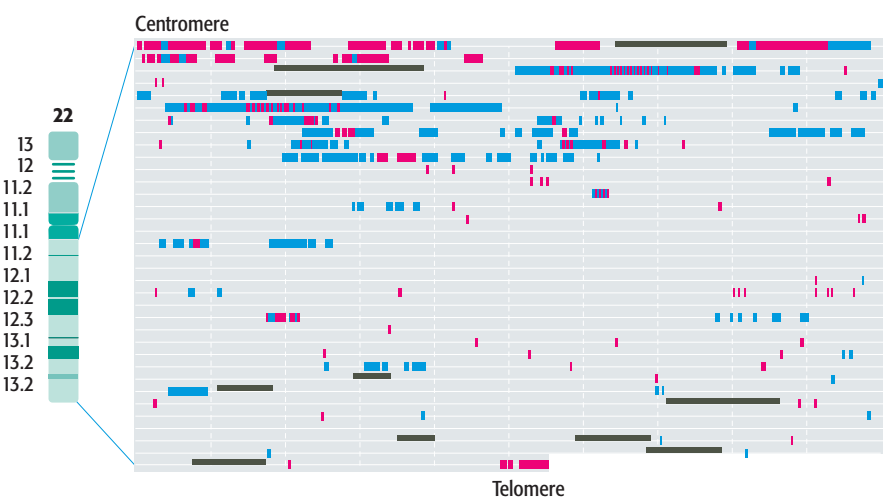


Figure 18.16 Segment duplications in the long arm of human chromosome 22. The diagram depicts the 34 Mb of the long arm of chromosome 22 as a series of thin horizontal lines, each one representing 1 Mb of DNA sequence, running from the centromere at the top to the telomere at the bottom. This analysis was based on the draft human genome sequence which contained eleven gaps in this region, shown as black bars. The pink boxes, which make up 3.9% of the 34 Mb, are sequences that are duplicated within this chromosome arm, and the blue boxes (6.4% of the total) are duplications of regions in other chromosomes.

are probably the products of relatively recent duplication events. Almost 200 segments, making up over 10% of this 34 Mb region of the human genome, appear to have arisen by duplications that occurred within the last 35 million years. Over 100 of these segments are duplicates of sequences present on other chromosomes, and the remainder are duplicates with both copies located within this arm.

The pattern of duplications in the long arm of chromosome 22 is fairly typical of the human genome as a whole. Individual duplications range from 1 to 400 kb in length and there is a distinct bias toward the areas adjacent to the centromeres, with relatively few duplications in the more distal regions of each chromosome arm. When considering the size of these segment duplications, we must bear in mind that the average human gene is 20–25 kb in length and that the genes are scattered sparsely throughout the genome. When the sequences of the segment duplications are examined it emerges that few of these duplication events result in duplication of entire genes, but that several involve parts of genes, and that some of these duplications lead to the upstream exons of one gene being placed alongside downstream exons of a second gene. Some of these new combinations are transcribed, but it is not clear whether the transcripts are functional. The evolutionary potential of segment duplications is therefore uncertain. What is clear, however, is that recombination between a pair of intrachromosomal duplications can result in deletion of the region between the duplications, and that deletions generated in this way can give rise to genetic disease. An example is Charcot–Marie–Tooth syndrome, which develops 5–15 years after birth and is characterized by degeneration of the peripheral nervous system, leading to weakness and difficulty in walking. The syndrome is caused by recombination between a pair of 24 kb segment duplications on human chromosome 17, deleting a 1.5 Mb segment of the genome. The loss of the genes contained in this segment gives rise to the disease.

Genome evolution also involves rearrangement of existing genes

The range of sizes observed for segment duplications in the human genome raises the possibility that the duplication events illustrated in Figure 18.11 could, as well as giving rise to new copies of genes, cause alterations *within* existing genes. This would be an alternative way in which novel protein functions could evolve. This is possible because most proteins are made up of structural domains, each comprising a segment of the polypeptide chain and hence encoded by a contiguous series of nucleotides (Figure 18.17). Rearrangement of domain-encoding gene segments could result in novel protein functions:

- **Domain duplication** would occur when the gene segment coding for a structural domain is duplicated by unequal crossing-over, replication slippage, or one of the other methods that we have considered for duplication of DNA sequences (Figure 18.18A). Duplication could result in the structural domain being repeated in the protein, which might itself be

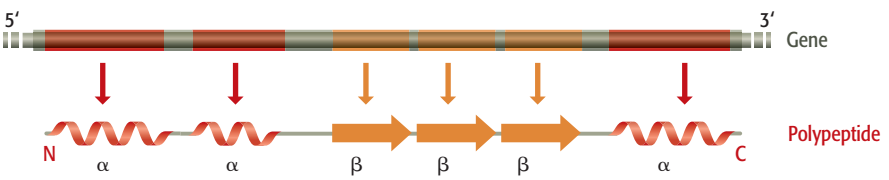


Figure 18.17 Each structural domain is an individual unit in a polypeptide chain and is coded by a contiguous series of nucleotides. In this simplified example, each secondary structure in the polypeptide is looked upon as an individual structural domain. In reality, most structural domains comprise two or more secondary structural units.

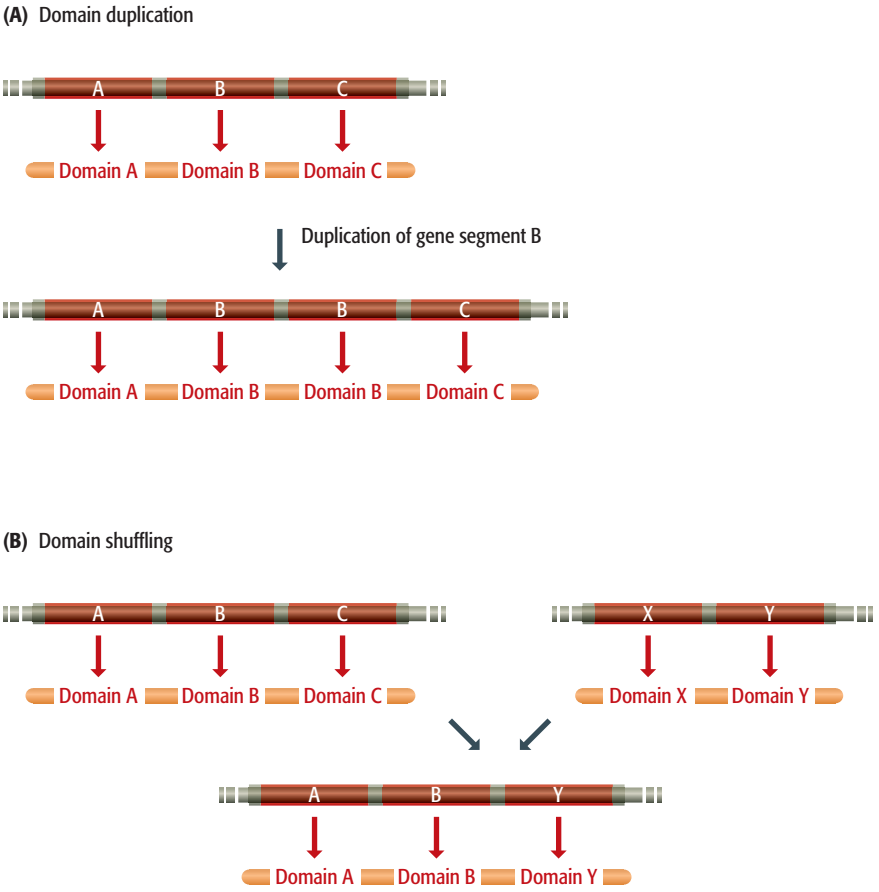


Figure 18.18 Creating new genes by (A) domain duplication and (B) domain shuffling.

advantageous, for example by making the protein more stable. The duplicated domain might also change over time as its coding sequence becomes mutated, leading to a modified structure that might provide the protein with a new activity. Note that domain duplication causes the gene to become longer. Gene elongation appears to be a general consequence of genome evolution, the genes of higher eukaryotes being longer, on average, than those of lower organisms.

- **Domain shuffling** would occur when segments coding for structural domains from completely different genes are joined together to form a new coding sequence that specifies a hybrid or mosaic protein, one that would have a novel combination of structural features and might provide the cell with an entirely new biochemical function (Figure 18.18B).

Implicit in these models of domain duplication and shuffling is the need for the relevant gene segments to be separated so that they can themselves be rearranged and shuffled. This requirement has led to the attractive suggestion that exons might code for structural domains. With some proteins, duplication or shuffling of exons does seem to have resulted in the structures seen today. An example is provided by the α2 Type I collagen gene of vertebrates, which codes for one of the three polypeptide chains of collagen. Each of the three collagen polypeptides has a highly repetitive sequence made up of repeats of the tripeptide glycine–X–Y, where X is usually proline and Y is usually hydroxyproline (Figure 18.19). The chicken α2 Type I gene is split into 52 exons, 42 of which cover the part of the gene coding for the glycine–X–Y repeats. Within

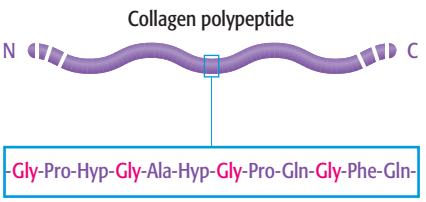


Figure 18.19 **The α2 Type I collagen polypeptide has a repetitive sequence described as Gly–X–Y.** Every third amino acid is glycine, X is often proline, and Y is often hydroxyproline (Hyp). Hydroxyproline is synthesized by posttranslational modification of proline (Section 13.3.3). The collagen polypeptide has a helical conformation, but one that is more extended than the standard α-helix.

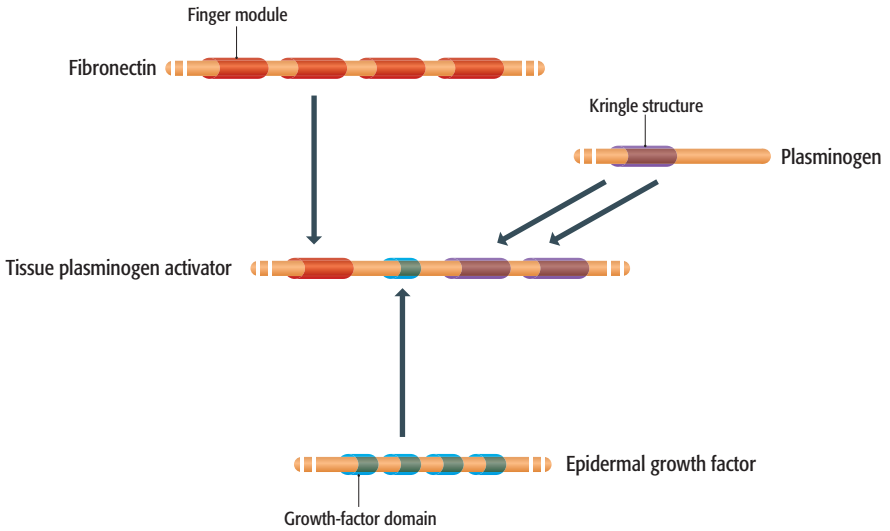


Figure 18.20 The modular structure of the tissue plasminogen activator gene.

this region, each exon encodes a set of complete tripeptide repeats. The number of repeats per exon varies but is 5 (in 5 exons), 6 (23 exons), 11 (5 exons), 12 (8 exons), or 18 (1 exon). Clearly this gene could have evolved by duplication of exons leading to repetition of the structural domains.

Domain shuffling is illustrated by tissue plasminogen activator (TPA), a protein found in the blood of vertebrates and which is involved in the blood clotting response. The TPA gene has four exons, each coding for a different structural domain (Figure 18.20). The upstream exon codes for a "finger" module that enables the TPA protein to bind to fibrin, a fibrous protein found in blood clots and which activates TPA. This exon appears to be derived from a second fibrin-binding protein, fibronectin, and is absent from the gene for a related protein, urokinase, which is not activated by fibrin. The second TPA exon specifies a growth factor domain which has apparently been obtained from the gene for epidermal growth factor and which may enable TPA to stimulate cell proliferation. The last two exons code for "kringle" structures which TPA uses to bind to fibrin clots; these kringle exons come from the plasminogen gene.

Type I collagen and TPA provide elegant examples of gene evolution but, unfortunately, the clear links that they display between structural domains and exons are exceptional and are rarely seen with other genes. Many other genes appear to have evolved by duplication and shuffling of segments, but in these the structural domains are coded by segments of genes that do not coincide with individual exons or even groups of exons. Domain duplication and shuffling still occur, but presumably in a less precise manner and with many of the rearranged genes having no useful function. Despite being haphazard, the process clearly works, as indicated by, among other examples, the number of proteins that share the same DNA-binding motifs (Section 11.1.1). Several of these motifs probably evolved *de novo* on more than one occasion, but it is clear that in many cases the nucleotide sequence coding for the motif has been transferred to a variety of different genes.

One possible mechanism for moving gene segments around a genome is in association with transposable elements. The transposition of a LINE-1 element (Section 9.2.1) can occasionally result in a short piece of the adjacent DNA

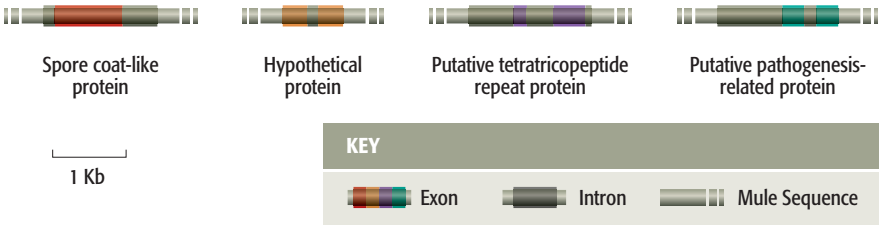


Figure 18.21 **MULEs often contain segments of genes captured from the host chromosomes.** The diagram shows five adjacent MULEs from chromosome 1 of rice. Colored segments are gene exons and gray segments between and adjacent to exons are introns. The thin gray bars at the ends of the structures are the terminal inverted repeats of the MULE transposons. The identities of the gene segments are given when known.

being transferred along with the transposon, a process called **3′ transduction**, the transferred segment being located at the 3′ end of the element. LINE-1 elements are sometimes found in introns so 3′ transduction could conceivably move downstream exons to new sites in a genome. Movement of exons and other gene segments might also be brought about by DNA transposons called *Mutator*-like transposable elements (MULEs), which are found in many eukaryotes but are especially common in plants. MULEs often contain within their DNA sequence segments of genes captured from the host genome (Figure 18.21). Transposition of a MULE would therefore move the captured segments to a new location. MULEs can collect segments of different genes as they travel around a genome, assembling new hybrid genes as they go. MULEs therefore provide an attractive way of driving gene evolution, but there are still several unanswered questions about their impact. In particular, it is not yet clear how frequently gene segments are able to escape from MULEs.

18.2.2 Acquisition of new genes from other species

The second possible way in which a genome can acquire new genes is to obtain them from another species. Comparisons of bacterial and archaeal genome sequences suggest that **lateral gene transfer** has been a major event in the evolution of prokaryotic genomes (Section 8.2.3). The genomes of most bacteria and archaea contain at least a few hundred kilobases of DNA, representing tens of genes, that appears to have been acquired from a second prokaryote.

There are several mechanisms by which genes can be transferred between prokaryotes but it is difficult to be sure how important these various processes have been in shaping the genomes of these organisms. Conjugation (Section 3.2.4), for example, enables plasmids to move between bacteria and frequently results in the acquisition of new gene functions by the recipients. On a day-to-day basis, plasmid transfer is important because it is the means by which genes for resistance to antibiotics such as chloramphenicol, kanamycin, and streptomycin spread through bacterial populations and across species barriers, but its evolutionary relevance is questionable. It is true that the genes transferred by conjugation can become integrated into the recipient bacterium's genome, but usually the genes are carried by composite transposons (see Figure 9.17B), which means that the integration is reversible and so might not result in a permanent change to the genome. A second process for DNA transfer between prokaryotes, transformation (Section 3.2.4), is more likely to have had an influence on genome evolution. Only a few bacteria, notably members of the *Bacillus*, *Pseudomonas*, and *Streptococcus* genera, have efficient mechanisms for the uptake of DNA from the surrounding environment, but efficiency of DNA uptake is probably not relevant when we are dealing with an evolutionary timescale. More important is the fact that gene flow by transformation can occur between any pair of prokaryotes, not just closely related ones (as is the case with conjugation),

and so could account for the transfers that appear to have occurred between bacterial and archaeal genomes (Section 8.2.3).

In plants, new genes can be acquired by polyploidization. We have already seen how autopolyploidization can result in genome duplication in plants (see Figure 18.13). **Allopolyploidy**, which results from interbreeding between two different species, is also common and, like autopolyploidy, can result in a viable hybrid. Usually, the two species that form the allopolyploid are closely related and have many genes in common, but each parent will possess a few novel genes or at least distinctive alleles of shared genes. For example, the bread wheat, *Triticum aestivum*, is a hexaploid that arose by allopolyploidization between cultivated emmer wheat, *T. turgidum*, which is a tetraploid, and a diploid wild grass, *Aegilops squarrosa*. The wild grass nucleus contained novel alleles for the high-molecular-weight glutenin genes which, when combined with the glutenin alleles already present in emmer wheat, resulted in the superior properties for breadmaking displayed by the hexaploid wheats. Allopolyploidization can therefore be looked upon as a combination of genome duplication and interspecies gene transfer.

Among animals, the species barriers are less easy to cross and it is difficult to find clear evidence for lateral gene transfer of any kind. Several eukaryotic genes have features associated with archaeal or bacterial sequences, but rather than being the result of lateral gene transfer, these similarities are thought to result from conservation during millions of years of parallel evolution. Most proposals for gene transfer between animal species center on retroviruses and transposable elements. Transfer of retroviruses between animal species is well documented, as is their ability to carry animal genes between individuals of the same species, suggesting that they might be possible mediators of lateral gene transfer. The same could be true of transposable elements such as P elements, which are known to spread from one *Drosophila* species to another, and *mariner*, which has also been shown to transfer between *Drosophila* species and which may have crossed from other species into humans.

18.3 Noncoding DNA and Genome Evolution

So far we have concentrated our attention on the evolution of the coding component of the genome. As coding DNA makes up only 1.5% of the human genome, our view of genome evolution would be very incomplete if we did not devote some time to considering noncoding DNA. The presence of extensive amounts of noncoding DNA in eukaryotic genomes is a puzzle for molecular evolutionists. Why is this apparently superfluous DNA tolerated? One possibility is that the noncoding DNA has a function that has not yet been identified and, as such, must be maintained because without it the cell would be nonviable. Plausible functions are not as difficult to identify as might be imagined. In several places in the preceding chapters the importance of chromatin structure has been stressed, including the attachment of chromatin to sites within the nucleus. Possibly, some of the noncoding component of a genome is involved in these aspects of genome organization. Alternatively, noncoding DNA might have a broad-ranging control function that so far has eluded discovery by molecular biologists. A final possibility is that noncoding DNA has no function but is tolerated by a genome because there is no selective pressure

to get rid of it. If this view is correct then possession of noncoding DNA is neither an advantage nor a disadvantage and so the noncoding DNA is simply propagated along with the coding DNA. According to this hypothesis, the noncoding DNA could simply be "junk" or might be parasitic "selfish DNA."

There is little that can be said about the evolution of much of the noncoding component of a genome. We envisage that duplications and other rearrangements have occurred through recombination and replication slippage, and that sequences have diverged through accumulation of mutations unfettered by the restraining selective forces acting on functional regions of the genome. We recognize that some parts of the noncoding DNA, for example the regulatory regions upstream of genes, have important functions, but for many other parts of the noncoding DNA, all we can say is that it evolves in an apparently random fashion. But this randomness does not apply to all components of the noncoding DNA. In particular, transposable elements and introns have interesting evolutionary histories and are of general importance in genome evolution.

18.3.1 Transposable elements and genome evolution

Transposable elements have a number of effects on evolution of the genome as a whole. The most significant of these is the ability of transposons to initiate recombination events that lead to genome rearrangements. This has nothing to do with the transposable activity of these elements, it simply relates to the fact that different copies of the same element have similar sequences and can therefore initiate recombination between two parts of the same chromosome or between different chromosomes (Figure 18.22). In many cases, the resulting rearrangement will be harmful because important genes will be deleted, but some instances where the result has been beneficial have been documented. Recombination between a pair of LINE-1 elements (Section 9.2.1) approximately 35 million years ago is thought to be the cause of the β-globin gene duplication that resulted in the Gγ and Aγ members of this gene family (see Figure 18.9).

Movement of transposons from one site to another can also have an impact on genome evolution. The possible relocation of gene segments during transposition of LINE elements and MULEs was mentioned in Section 18.2.1. Transposition has also been associated with altered patterns of gene expression. For example, the efficiency with which DNA-binding proteins that are attached to upstream regulatory sequences can activate transcription of a gene might be affected if a transposon moves into a new site immediately upstream of the gene (Figure 18.23). The presence of promoters and/or enhancers within the transposon might also affect transcription by subjecting the adjacent gene to an entirely new regulatory regime. An interesting example of transposon-directed gene expression occurs with the mouse gene *Slp*, which codes for a protein involved in the immune response. The tissue specificity of *Slp* is conferred by an enhancer located within an adjacent retrotransposon. There are also examples where insertion of a transposon into a gene has resulted in an altered splicing pattern.

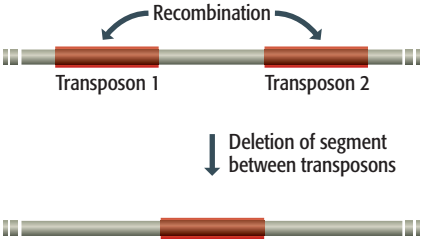


Figure 18.22 Recombination between pairs of repeated sequences, such as transposons, can result in deletion of segments of the genome.

18.3.2 The origins of introns

Ever since introns were first discovered in the 1970s their origins have been debated. There are few controversies surrounding the Group I, II, and III types (see Table 12.2) as it is generally accepted that all these self-splicing

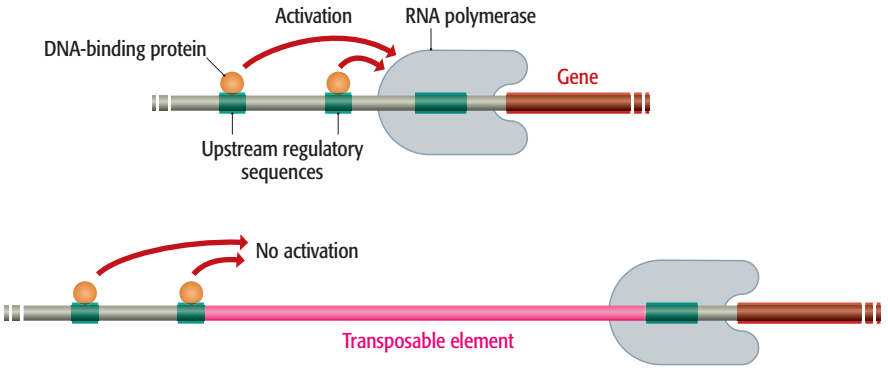


Figure 18.23 Insertion of a transposon into the region upstream of a gene could affect the ability of DNA-binding proteins to activate transcription.

introns evolved in the RNA world and have survived ever since without undergoing a great deal of change. The problems surround the origin of the GU–AG introns, the ones that are found in large numbers in eukaryotic nuclear genomes.

"Introns early" and "introns late": two competing hypotheses

A number of proposals for the origin of GU–AG introns have been put forward but the debate is generally considered to be between two opposing hypotheses:

- **"Introns early"** states that introns are very ancient and are gradually being lost from eukaryotic genomes.
- **"Introns late"** states that introns evolved relatively recently and are gradually accumulating in eukaryotic genomes.

There are several different models for each hypothesis. For "introns early" the most persuasive model is the one also called the "exon theory of genes," which holds that introns were formed when the first DNA genomes were constructed, soon after the end of the RNA world. These genomes would have contained many short genes, each derived from a single coding RNA molecule and each specifying a very small polypeptide, perhaps just a single structural domain. These polypeptides would probably have had to associate together into larger multidomain proteins in order to produce enzymes with specific and efficient catalytic mechanisms (Figure 18.24). To aid the synthesis of a multidomain enzyme it would have been beneficial for the enzyme's individual polypeptides to become linked into a single protein, such as we see today. It is envisaged that this was achieved by splicing together the transcripts of the relevant minigenes, a process that was aided by rearranging the genome so that groups of minigenes specifying the different parts of individual multidomain proteins were positioned next to each other. In other words, the minigenes became exons and the DNA sequences between them became introns.

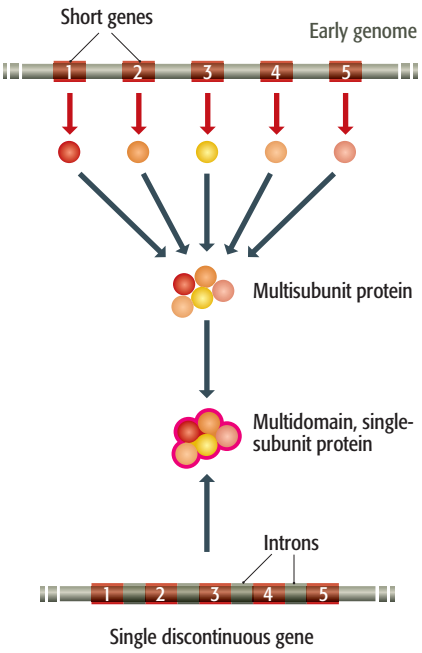


Figure 18.24 The "exon theory of genes." The short genes of the first genomes probably coded for single-domain polypeptides that would have had to associate together to form a multisubunit protein to produce an effective enzyme. Later the synthesis of this enzyme could have been made more efficient by linking the short genes together into one discontinuous gene coding for a multidomain single-subunit protein.

According to the exon theory of genes and other "introns early" hypotheses, all genomes originally possessed introns. But we know that bacterial genomes do not have GU–AG introns, so if these hypotheses are correct then we must assume that for some reason introns became lost from the ancestral bacterial genome at an early stage in its evolution. This is a stumbling block because it is difficult to envisage how a large number of introns could be lost from a genome without risking the disruption of many gene functions. If an intron is removed from a gene with any imprecision then a part of the coding region will be lost or a frameshift mutation will occur, both of which would be expected to inactivate the gene. The "introns late" hypothesis avoids this

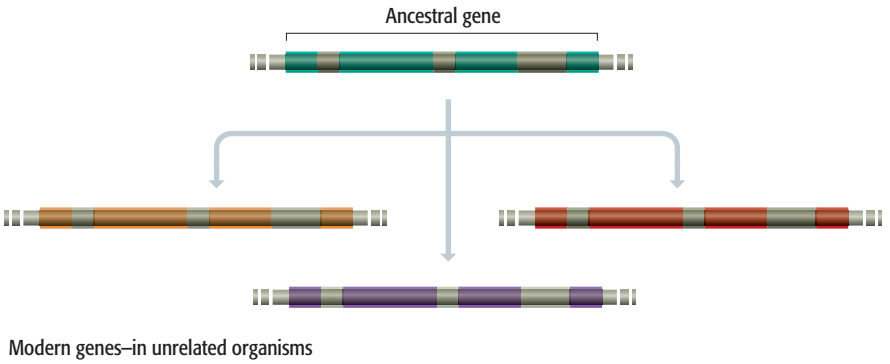


Figure 18.25 One prediction of the "introns early" hypothesis is that the positions of introns in homologous genes should be similar in unrelated organisms, because all these genes are descended from an ancestral intron-containing gene.

problem by proposing that, to begin with, no genes had introns, these structures invading the early eukaryotic nuclear genome and subsequently proliferating into the numbers seen today. The similarities between the splicing pathways for GU–AG and Group II introns (Section 12.2.4) suggest that the invaders that gave rise to GU–AG introns might well have been Group II sequences that escaped from organelle genomes. However, the similarity between GU–AG and Group II introns does not prove the "introns late" view, because it is equally possible to devise an "introns early" model, different to the exon theory of genes, in which Group II sequences gave rise to GU–AG introns, but at a very early stage in genome evolution.

The current evidence disproves neither hypothesis

One of the reasons why the debate regarding the origin of GU–AG introns has continued for over 25 years is because evidence in support of either hypothesis has been difficult to obtain and is often ambiguous. One prediction of "introns early" is that there should be a close similarity between the positions of introns in homologous genes from unrelated organisms, because all these genes are descended from an ancestral intron-containing gene (Figure 18.25). Early support for "introns early" came when this was shown to be the case for four introns in animal and plant genes for triosephosphate isomerase. However, when a larger number of species was examined, the positions of the introns in this gene became less easy to interpret: it appeared that introns had been lost in some lineages but gained in others. This scenario fits both "introns early" and "introns late," as both hypotheses allow for the loss, gain, or repositioning of introns by recombination events occurring in individual lineages. When many genes in many organisms are examined, the general picture that emerges is that intron numbers have gradually increased during the evolution of animal genomes, this being put forward as evidence for "introns late," despite the fact that animal mitochondrial genomes do not contain Group II introns that could supplement the existing nuclear introns by repeated invasions. Intron numbers must therefore have increased by recombination events.

An alternative approach has been to try to correlate exons with protein structural domains, as the "introns early" hypothesis predicts that such a link should be evident, even allowing for the fuzzying effects of evolution since the primitive minigenes were assembled into the first real genes. Again, the first evidence to be obtained supported "introns early." A study of vertebrate globin proteins concluded that each of these comprises four structural domains, the first corresponding to exon 1 of the globin gene, the second and third to exon 2, and the fourth to exon 3 (Figure 18.26). The prediction that there should be globin genes with another intron that splits the second and

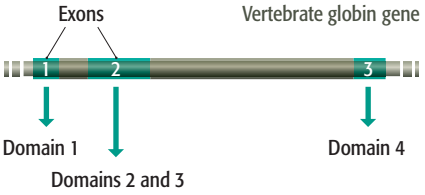


Figure 18.26 A vertebrate globin gene showing the relationship between the three exons and the four domains of the globin protein.

third domains was found to be correct when the leghemoglobin gene of soybean was shown to have an intron at exactly the expected position. Unfortunately, as more globin genes were sequenced, more introns were discovered—more than ten in all. The positions of the majority of these do not correspond to junctions between domains.

The globin genes therefore conform with the general principle that emerged from our discussion of domain shuffling (Section 18.2.1): that in most cases there are no clear links between gene exons and protein structural domains. But is our definition of "structural domain" correct? A structural domain within a protein may not simply correspond with a group of secondary structures such as α-helices and β-sheets. A more subtle interpretation might be that a structural domain is a polypeptide segment whose amino acids are less than a certain distance apart in the protein's tertiary structure. It has been suggested that when this definition is adopted there is a better correlation between structural domain and exons.

18.4 The Human Genome: the Last Five Million Years

Although the evolutionary history of humans is controversial, it is generally accepted that our closest relative among the primates is the chimpanzee and that the most recent ancestor that we share with the chimpanzee lived 4.6–5.0 million years ago. Since the split, the human lineage has embraced two genera—*Australopithecus* and *Homo*—and a number of species, not all of which were on the direct line of descent to *Homo sapiens* (Figure 18.27). The result is us, a novel species in possession of what are, at least to our eyes, important biological attributes that make us very different from all other animals. So how different are we from the chimpanzees?

As far as our genomes are concerned the answer is 1.73%, this being the extent of the nucleotide sequence dissimilarity between humans and chimpanzees. Indeed, when the human and chimpanzee genomes are compared it is much easier to find similarities than differences. The degree of nucleotide sequence identity within the coding DNA is greater than 98.5%, with 29% of

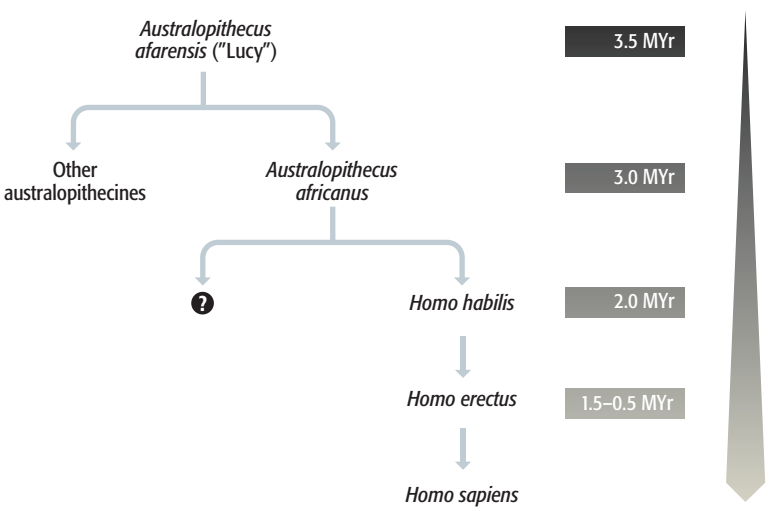


Figure 18.27 One possible scheme for the evolution of modern humans from australopithecine ancestors. There are many controversies in this area of research and several different hypotheses have been proposed for the evolutionary relationships between different fossils. Abbreviation: MYr, million years.

the genes in the human genome coding for proteins whose amino acid sequences are identical to the sequences of their counterparts in chimpanzees. Even in the noncoding regions of the human and chimpanzee genomes the nucleotide identity is rarely less than 97%. Gene order is almost the same in the two genomes, and the chromosomes have very similar appearances. At this level, the most dramatic difference is that human chromosome 2 is two separate chromosomes in chimpanzees (Figure 18.28), so chimpanzees, as well as other apes, have 24 pairs of chromosomes whereas humans have just 23 pairs. The alphoid DNA sequences present at human centromeres (Section 7.1.2) are quite different from the equivalent sequences in chimpanzee and gorilla chromosomes, and Alu elements (Section 9.2.1) are more prevalent in the human genome, but these features probably tell us more about the evolution of repetitive DNA than about the differences between humans and chimpanzees.

Comparisons of the human and chimpanzee genomes have also failed to reveal changes to individual genes which might somehow be key to the special attributes of humans. Analyses designed to reveal genes that have been under positive selection in the human lineage have identified several associated with amino acid breakdown, in line with the greater proportion of meat eaten by humans compared with chimpanzees. Genes providing protection against human diseases such as tuberculosis and malaria have also been subject to positive selection. But no genes with clear roles in brain or neuronal development have been uncovered by this type of analysis. The only substantial difference in gene structure is that humans lack a 92 bp segment of the gene for *N*-glycolyl-neuraminic acid hydroxylase and so cannot synthesize the hydroxylated form of *N*-glycolyl-neuraminic acid, which is present on the surfaces of some chimpanzee cells. This may have an effect on the ability of certain pathogens to enter human cells, and could possibly influence some types of cell–cell interaction, but the difference is not thought to be particularly significant. Two other genes that are functional in chimpanzees appear to have been inactivated by point mutations in the human genome, one of these coding for a T-cell receptor and the other for a hair keratin protein, but neither of these changes is likely to have had any substantial impact on human evolution. In any case, it seems impossible that the special features of humans could have arisen as a result of a gene knockout. Instead we need to find genes whose activities have changed rather than been lost. In this regard, considerable interest has centered on the gene for the FOXP2 transcription factor, as defects in this protein result in the human disability called dysarthria, characterized by a difficulty in articulating speech. This gene might therefore underlie the human ability for language. There are indeed two amino acid differences between the FOXP2 genes in humans and chimpanzees, suggesting that this gene has undergone relatively recent change, but a direct link between these amino acid differences and human linguistic ability is elusive.

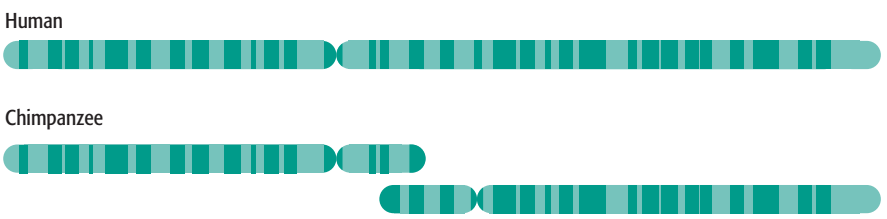


Figure 18.28 Human chromosome 2 is a fusion of two chromosomes that are separate in chimpanzees.

It is now becoming clear that many, if not all, of the key differences between humans and chimpanzees are likely to lie not with the genomes themselves but with the way the genomes are expressed. Attention is therefore moving from the genomes to the transcriptomes and proteomes, and these studies are beginning to suggest that the pattern of genome expression in the brain has undergone significant change in the human lineage since the divergence of humans and chimpanzees. Of course, this is precisely what we might expect, as it is clearly our brains that distinguish us from chimpanzees and other animals. The key question, which has not yet been answered, is whether the identities of the genes that are up- or downregulated in the human brain are informative in any way.

Summary

It is thought that the first polynucleotides to evolve, several billion years ago, were made of RNA rather than DNA. These RNA molecules probably combined a self-replicating ability with some enzymatic activity, and it is possible that enclosure of these in simple, lipid envelopes gave rise to the progenitors of the first cells. Experiments have suggested that short peptides might be constructed by catalytic RNAs, these peptides taking over some of the enzymatic functions of the ribozymes. DNA probably evolved as a more stable version of the RNA protogenomes. During evolution, there have been at least two periods when more complex genomes have emerged, in each case ones with increased gene numbers. Gene duplication is known to be an important event that can result in a genome acquiring new genes. The globin gene superfamily arose by a series of gene duplications whose pattern and timing can be inferred by making comparisons between the sequences of the globin genes in existence today. Duplication events have also played an important role during evolution of the homeotic selector genes that specify the body plan of eukaryotes. Duplication of entire genomes is also possible and is thought to have occurred in the lineages leading to *Saccharomyces cerevisiae*, to *Arabidopsis thaliana*, and also to vertebrates. Smaller duplications, of a few tens of kilobases, have occurred regularly in the recent evolution of the human genome. Some of these have resulted in new combinations of exons, and some genes have clearly arisen by duplication or shuffling of protein domains coded by individual exons. Exons can also be transported around a genome by attachment to transposable elements. Lateral gene transfer results in the acquisition of genes from other species. This has been a regular event in the evolution of prokaryotic genomes but is much less common in eukaryotes, except possibly in plants which can form new polyploids by fusion of gametes from related species. The origin of introns is unclear, with evidence supporting both the "introns early" and "introns late" hypotheses. The former suggests that the first genomes had introns which are gradually being lost, and the latter proposes there has been a gradual increase in intron numbers. Five million years ago the human and chimpanzee lineages diverged. The genomes of humans and chimpanzees still display 98.3% nucleotide sequence identity, with many genes giving rise to identical protein products. Identifying specific features of the human genome that make us human is proving difficult, and it is now thought that the most important differences between humans and chimpanzees might not be the genomes themselves but the way in which the genomes are expressed.

Multiple Choice Questions

*Answers to odd-numbered questions can be found in the Appendix

18.1* The first biochemical systems on Earth were probably centered on which types of biomolecule?

a. Carbohydrates.
b. DNA.
c. Proteins.
d. RNA.

18.2. The term protogenome describes:

a. The first DNA genomes.
b. The first cellular RNA genomes.
c. Early RNA molecules that could self-replicate and direct biochemical reactions.
d. The first polymeric RNA molecules.

18.3* Which of the following statements about the transition from RNA to DNA genomes is FALSE?

a. The phosphodiester bonds in DNA are more stable than those in RNA.
b. RNA was readily oxidized by the oxygen in the Earth's atmosphere to form DNA.
c. The replacement of uracil with thymine provided more stability to DNA.
d. Double-stranded DNA provided a mechanism for repairing genetic material.

18.4. What is concerted evolution?

a. A process by which two gene products evolve to interact with each other.
b. A process by which genes are duplicated to provide additional gene products.
c. A process by which genes are mutated so that they can be recruited to new gene families.
d. A process by which the members of a gene family retain the same or similar nucleotide sequences.

18.5* Which of the following processes is thought to underlie concerted evolution?

a. Gene conversion.
b. Lateral gene transfer.
c. Programmed mutation.
d. Transposition.

18.6. How can the pattern of gene duplications within a multigene family be determined?

a. By comparing the nucleotide sequences of the genes.
b. By comparing the physiological functions of the gene products.
c. By comparing the structures of the gene products.
d. By comparing the locations of the genes in the genome.

18.7* Which of the following processes, resulting in gene duplication, occurs when DNA is exchanged between a pair of chromatids within a single chromosome?

a. DNA amplification.
b. Replication slippage.
c. Unequal crossing-over.
d. Unequal sister chromatid exchange.

18.8. What is the result of autopolyploidy?

a. A nucleus derived from the fusion of gametes from two different species.
b. A nucleus containing extra copies of a single chromosome.
c. A nucleus derived from the fusion of two diploid gametes from the same species.
d. A nucleus that contains extra copies of the sex chromosomes.

18.9* Which of the following events could lead to the evolution of a new gene that contains exons from two or more other genes?

a. Domain duplication.
b. Domain shuffling.
c. Gene conversion.
d. Gene duplication.

18.10. Which process describes the transposition of a DNA segment along with an adjacent transposon?

a. 3′ transduction.
b. Gene conversion.
c. Retrotransposition.
d. Transformation.

18.11* Within prokaryotes, lateral gene transfer is most likely to occur as a result of which of the following processes?

a. Conjugation.
b. Transduction.
c. Transformation.
d. Transposition.

18.12. The splicing pathway for GU–AG introns is most similar to the self-splicing pathway of which group of introns?

a. Group I.
b. Group II.

continued …

Multiple Choice Questions (continued)

*Answers to odd-numbered questions can be found in the Appendix

c. Group III.
d. None of the above.

18.13.* The differences between humans and chimpanzees are most likely due to:
a. The presence of extra genes in the human genome.
b. The deletion of chimpanzee genes from the human genome.
c. Changes in the amino acid sequences of proteins involved in speech.
d. Differences between the expression patterns of the two genomes.

Short Answer Questions

*Answers to odd-numbered questions can be found in the Appendix

18.1.* How does the atmosphere of Earth today differ from the atmosphere present when life first evolved?

18.2. How might some of the amino acids, nucleotide bases, and sugars have been synthesized before life evolved?

18.3.* If all present-day living organisms evolved from a single origin, is it possible or likely that there were other origins of biological systems on the ancient Earth? Explain your answer.

18.4. Provide a time line for the evolution of living organisms from the formation of the Earth to the appearance of the first hominoids.

18.5.* Which evolutionary periods are associated with increases in gene numbers in living organisms?

18.6. Discuss how the homeotic selector genes provide a good example of genome evolution by gene duplication.

18.7.* What are the characteristics of retrogenes and how do these arise in the genome?

18.8. What is the evidence for whole genome duplication in the evolutionary lineage leading to *Saccharomyces cerevisiae*?

18.9.* How can segment duplications give rise to new genes?

18.10. In which organisms is allopolyploidy likely to occur?

18.11.* What is the "exon theory of genes"?

18.12. Why do humans and chimpanzees have different numbers of chromosomes?

In-depth Problems

*Guidance to odd-numbered questions can be found in the Appendix

18.1.* Give the evidence for the conclusion that the *Arabidopsis thaliana* genome underwent four rounds of genome duplication between 100 million and 200 million years ago.

18.2. Are the examples of domain duplication and domain shuffling given in Section 18.2.1 special cases or are they representative of genome evolution in general?

18.3.* One of the initial publications of the draft human genome sequence (IHGSC [International Human Genome Sequencing Consortium] [2001] Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921) suggested that between 113 and 223 human genes might have been acquired from bacteria by lateral gene transfer. Subsequently it was concluded that this interpretation is incorrect and these genes are not bacterial in origin. What was the evidence that supported lateral transfer of these genes and why was this evidence subsequently discounted?

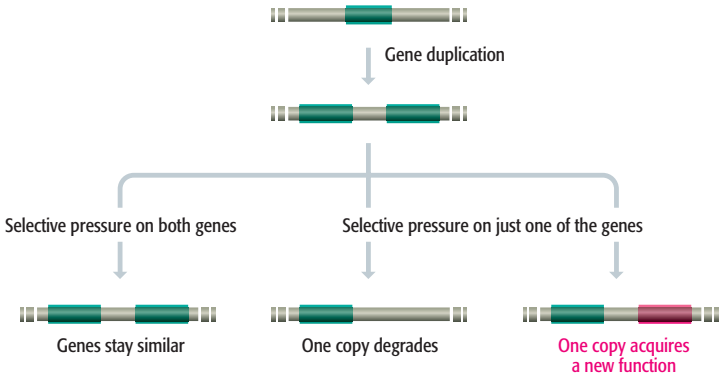
18.4. Evaluate the "introns early" and "introns late" hypotheses.

18.5.* To what extent do you believe it will be possible to determine the genetic basis to the special attributes of humans from comparisons between the genome sequences of humans and other primates?

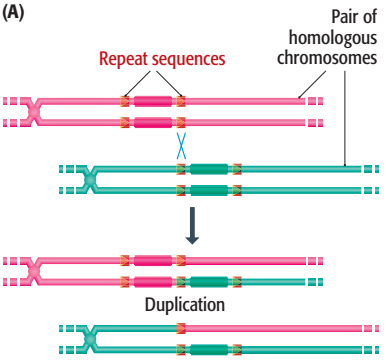
Figure Tests

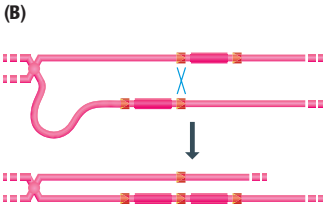
*Answers to odd-numbered questions can be found in the Appendix

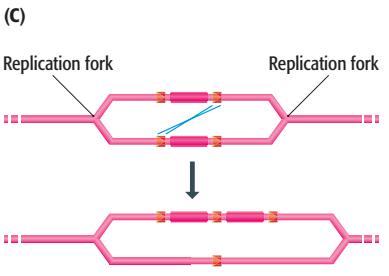
18.1* Discuss how new genes can arise from gene duplication, based on the pathways shown in the figure. Which of the pathways would give rise to a pseudogene?



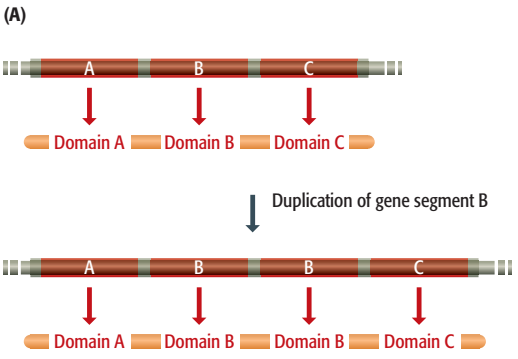
18.2. Which of the following models, (A), (B), or (C), represents gene duplication arising by unequal sister chromatid exchange?

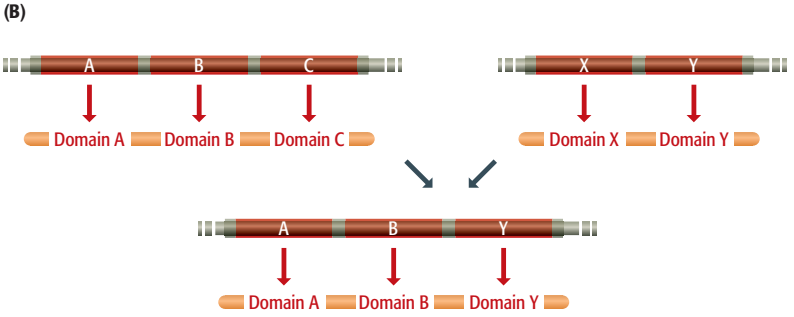






18.3* Discuss how the processes shown in the figures result in evolution of new genes.



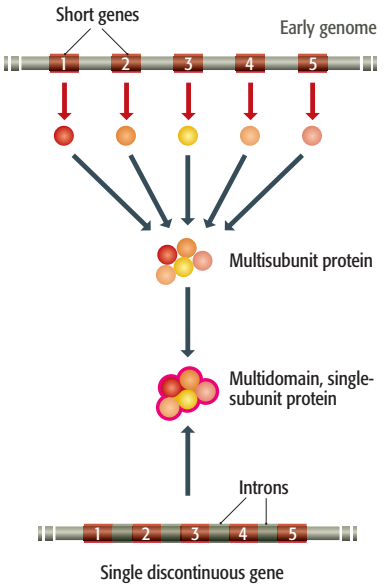


continued ...

Figure Tests (continued)

*Answers to odd-numbered questions can be found in the Appendix

18.4. Are the evolutionary events depicted in this figure supportive of the "introns early" or "introns late" hypothesis? What is represented by this figure?



Further Reading

Landmark books

Maynard Smith, J. and Szathmáry, E. (1995) *The Major Transitions in Evolution*. WH Freeman, Oxford. *Begins with the origin of life and ends with the evolution of human language.*

Ohno, S. (1970) *Evolution by Gene Duplication*. George Allen and Unwin, London.

The RNA world and the origins of genomes

Bartel, D.P. and Unrau, P.J. (1999) Constructing an RNA world. *Trends Genet.* **12:** M9–M13.

Freeland, S.J., Knight, R.D. and Landweber, L.F. (1999) Do proteins predate DNA? *Science* **286:** 690–692.

Lohse, P.A. and Szostak, J.W. (1996) Ribozyme-catalysed amino-acid transfer reactions. *Nature* **381:** 442–444.

Miller, S.L. (1953) A production of amino acids under possible primitive Earth conditions. *Science* **117:** 528–529.

Orgel, L.E. (2000) A simpler nucleic acid. *Science* **290:** 1306–1307. *Pyranosyl RNA.*

Robertson, M.P. and Ellington, A.D. (1998) How to make a nucleotide. *Nature* **395:** 223–225.

Unrau, P.J. and Bartel, D.P. (1998) RNA-catalysed nucleotide synthesis. *Nature* **395:** 260–263.

Gene duplications

Amores, A., Force, A., Yan, Y.-L., *et al.* (1998) Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282:** 1711–1714.

Wagner, A. (2001) Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends Genet.* **17:** 237–239.

Genome and segment duplications

Eichler, E.E. (2001) Recent duplication, domain accretion and dynamic mutation of the human genome. *Trends Genet.* **17:** 661–669.

Goffeau, A. (2004) Seeing double. *Nature* **430:** 25–26. *Comparisons between different yeast species indicate a genome duplication in the Saccharomyces cerevisiae lineage.*

Vision, T.J., Brown, D.G. and Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science* **290:** 2114–2117.

Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.

Transposable elements and genome evolution

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569–573.

Kazazian, H.H. (2000) L1 retrotransposons shape the mammalian genome. *Science* **289:** 1152–1153.

Origins of introns

de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W. and Gilbert, W. (1996) Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl Acad. Sci. USA* **93:** 14632–14636.

Gilbert, W. (1987) The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* **52:** 901–905.

Gilbert, W., Marchionni, M. and McKnight, G. (1986) On the antiquity of introns. *Cell* **46:** 151–153.

Palmer, J.D. and Logsdon, J.M. (1991) The recent origin of introns. *Curr. Opin. Genet. Dev.* **1:** 470–477.

Humans and chimpanzees

Balter, M. (2005) Are human brains still evolving? Brain genes show signs of selection. *Science* **309:** 1662–1663.

Chou, H.H., Takematsu, H., Diaz, S., *et al.* (1998) A mutation in human CMP-sialic acid hydroxylase occurred after the *Homo–Pan* divergence. *Proc. Natl Acad. Sci. USA* **95:** 11751–11756.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M. and Paabo, S. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309:** 1850–1854.

Li, W.H. and Saunders, M.A. (2005) The chimpanzee and us. *Nature* **437:** 50–51. *Describes the key differences between the human and chimpanzee genomes.*

Muchmore, E.A., Diaz, S. and Varki, A. (1998) A structural difference between the cell surfaces of humans and the great apes. *Am. J. Phys. Anthropol.* **107:** 187–198.

This page left intentionally blank to match
pagination with the printed text.